Leveraging the Positive Deviance Approach using Big Data

A thesis submitted to the University of Manchester for the degree of

Doctor of Philosophy in the Faculty of Humanities

2021

Basma Albanna

Global Development Institute

School of Environment, Education and Development

Table of Contents

Word count: 77, 963

List of Figures

List of Tables

List of Abbreviations

AGEB	Area Geo-Estadistica Basica
BAPPENAS	Ministry of National Development Planning of the Republic of Indonesia
BD	Big Data
BD4D	Big Data for Development
BDPD	Big Data-based Positive Deviance
CDRs	Call Detail Records
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station data
CMAP	CPC merged analysis of precipitation
DAAD	German Academic Exchange Service
DADs	Discovery and Action Dialogues
DOI	Diffusion of Innovations
DPPD	Data Powered Positive Deviance
EO	Earth Observation
ERASMUS+	European Region Action Scheme for the Mobility of University Students
EVI	Enhanced Vegetation Index
GPM	Global Precipitation Measures
GIZ	Deutsche Gesellschaft für Internationale Zusammenarbeit
HE	Homologues Environment
IaaS	Infrastructure as a service (IaaS)
IS	Information Systems

ISI	International Scientific Indexing
L1Q	Level One Questions
L2Q	Level Two Questions
LDA	Latent Dirichlet Allocation
ML	Machine Learning
MODIS	Moderate Resolution Imaging Spectroradiometer
MOOCs	Massive Open Online Courses
MSE	Means Squared Error
NDVI	Normalized Difference Vegetation Index
PCs	Principal Components
PCA	Principal Component Analysis
PD	Positive Deviance
PDI	Positive Deviance Inquiry
PDs	Positive Deviants
PLJ	Pulse Lab Jakarta
PLS	Partial Least Square
PODIS	Village Potential Survey
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analysis
RMSE	Root Mean Squared Error
SAVI	Soil-Adjusted Vegetation Index
SCF	Save the Children Foundation
SLR	Systematic Literature Review

UNGP	United Nations Global Pulse
UNDP	United Nations Development Program
VAWG	Violence Against Women and Girls

Abstract

This thesis presents a method that combines non-traditional digital data, namely big data, and traditional data to identify and characterise outperformers in development-related challenges. It builds on the "Positive Deviance" (PD) approach for development, which is based on the observation that in every population there are individuals or communities who, despite facing similar challenges and limitations, achieve better results than their peers. This approach focuses on these outliers (or positive deviants) in order to discover unusual practices and strategies that successfully solve complex problems – particularly where conventional solutions failed. In order to build this method, I first conducted a systematic literature review to identify the opportunities and challenges of using big data in the positive deviance approach and outlined a preliminary analytical framework that could guide the use of such data in the PD approach. Following that, I tested and validated this framework in multiple case studies which supported the iterative development and refinement of a method I refer to as "Data Powered Positive Deviance" (DPPD).

In the first case study, I used online data along with traditional data sources, such as surveys and interviews, to identify and characterise Egyptian information system researchers who were able to outperform their peers in publication outcomes. I applied the same framework during a six-month fellowship at the United Nations Pulse Lab Jakarta, to identify and validate positively deviant rice-growing villages in Indonesia using official statistics and administrative data along with earth observation big data. This fellowship was part of a global initiative collaboratively created by the GIZ data lab, UNDP accelerator labs, Pulse Lab Jakarta and the University of Manchester Centre for digital development. It builds and scales the DPPD method developed in this study, to see if and how we might use big data-based positive deviance to tackle development challenges. In addition to the Indonesia project, four action research projects were implemented as part of this initiative to identify and understand: farmers achieving higher than usual cereal crop productivity in Niger; cattle farmers in Ecuador who are deforesting below average rates; public spaces in Mexico City where women are safest; and communities in Somalia which are able to preserve their rangelands despite the frequent droughts. I was heavily involved in the implementation of those action research projects as I was the methodological and technical lead. This initiative enabled me to test and fine tune the method with real life development problems and practitioners in a much wider domain space.

The DPPD method, presented in this thesis, provides a tool for development professionals to identify outperformance in different development sectors by mixing analytical insights from traditional and non-traditional data. Such insights should help amplify innovative, locally sourced and evidence-informed solutions to development challenges.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

I. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

II. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

III. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

IV. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=2442 0), in any relevant Thesis restriction declarations deposited in the University Library, the University Library's regulations (see http://www.library.manchester.ac.uk/about/regulations/) and in the University's policy on Presentation of Theses

Acknowledgments

I shall forever be indebted to my primary supervisor, Professor Richard Heeks, whose support, wisdom, patience and example, as a teacher, supervisor, co-author and colleague, have been the greatest single influence on my progression in this PhD. I appreciate all that he has done to encourage my growth as both, a development researcher and practitioner and all the opportunities he put in my way that shaped my doctoral journey. My sincere gratitude to my second supervisor, Professor Julia Handl, who in quite different ways have provided essential guidance and critique at just the right time. I am extremely grateful that you both took me on as a student and continued to have faith in me over the years.

I am grateful for the comments of my internal independent reviewer Dr. Jaco Renken, who provided critical and constructive comments for three consecutive years. I am extremely grateful to the School of Environment, Education and Development for their generous studentship, providing me with the financial means to complete this PhD. I am also grateful to the Global Development Institute (GDI) family. It has been a privilege to be part of the GDI community with its considerate staff and students who together have provided a warm and inspiring environment. In particular, I want to thank Dr. Admos Chimhowu for his kindness and support.

I also want to thank the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) data lab, particularly Catherine Vogel, for being an early believer of the method presented in this thesis and a key player in founding the Data Powered Positive Deviance initiative which enabled me to test and experiment the method though a number of action research projects across the globe. I'm also thankful for the in-country United Nations Development Programme Accelerator Lab teams, the GIZ projects and all their local partners for implementing the Ecuador, Mexico, Niger and Somalia action research projects. I'm grateful for the United Nations Global Pulse and Pulse Lab Jakarta for their fellowship and their support in implementing the Indonesia project. I thank the many helpful colleagues from the DPPD initiative who have contributed to the development of the DPPD method with their valuable input and insights, specifically Andreas Pawelke, Andreas Gluecker, Jeremy Boy, Dharani Dhar Burra and Michael Dyer.

Thanks to Assem, my partner, for his love, support and patience throughout this PhD and for all the sacrifices he has made in order for me to pursue this degree. Finally, I am grateful for my parents whose constant love and unconditional support keep me motivated and confident. My accomplishments and success are because they believed in me. Deepest thanks to my two siblings, who keep me grounded and are always supportive of my dreams.

Chapter One: Introduction to the Study

Most development practitioners are still using the traditional need-based approaches to development, which are based on identifying problems to provide top-down, externally imposed solutions. Need-based approaches work well in addressing technical challenges that can be solved by the knowledge of experts. However, adaptive challenges (e.g. the need within a community for healthier eating or to increase awareness of the negative effects of child marriage), that require a behavioural change, obligate new learning that can only be found within the community (Singhal et al. 2010; Saïd Business School 2010; Pascale et al. 2010; Nel 2017). As a result, asset-based approaches came into existence, which capitalize on a community's inherent wisdom, assets and capabilities in solving their own problems. Positive deviance (PD) is one of the asset-based approaches that have been used across many disciplines since its proven success as a practical strategy by Sternin et al. (1997). It is based on the observation that in every community or organization, a few individuals or groups use uncommon practices and behaviours to achieve better solutions to problems than their peers who face the same challenges and barriers (Pascale, Sternin & Sternin 2010). Those individuals are referred to as positive deviants and adopting their solutions is referred to as the PD approach.

This chapter aims at introducing the PD approach, its promise for development and its current limitations, while demonstrating the potential value of using big data in this approach to leverage its benefits and mitigate some of its challenges. The chapter is structured as follows: Sections 1.1 and 1.2 introduce the concept of PD and how it evolved into a practical strategy for development with evident benefits in a variety of applications, followed by an overview of the conceptual and practical challenges of the current PD approach. Section 1.3 provides a brief introduction to big data, its characteristics and the benefits of applying it in development. Section 1.4 proposes the possibility of a big data-based PD method that combines the use of big data and PD while presenting some of the foreseen opportunities. Section 1.5 poses the primary research question and outlines research objectives before concluding with the methodological strategy. Finally, section 1.6 presents the thesis structure and format.

1.1 The Concept of Positive Deviance

Strong affinities exist between PD and many theories in the literature. Since the early 1900s the study of deviance has been addressed extensively in the sociology literature but with a clear focus on negative deviance. It was not until 1964 that the study of PD came into place. Four main views on PD where provided by the sociology literature: statistical, superconformity, reactivist and normative (Spreitzer & Sonenshein 2004). The statistical view uses a bell curve to represent social behaviour, having positive deviance at the right end, negative deviance at the left and conforming behaviours in the middle (Wolfgang 1965). Superconformity views PD as excessive conformity to norms i.e. behaviours exceeding the agreed-upon limits of norms (Dodge 1985; Ewald and Jiobu 1985; Hughes and Coakley 1991). It is conceptualized as pronormative as long as it does not extend beyond the limits considered appropriate by the referent group. The reactivist (or subjectivist) view refers to PD as behaviours that are labelled and evaluated in a positive way (Dodge 1985). It requires both the observation of the deviant behaviour and labelling this behaviour as positive. The normative view defines PD as a departure from the norm in a constructive way (Warren 2003). It was defined by Spreitzer and Sonenshein (2003) as "intentional behaviours that depart from the norms of a referent group in honourable ways". They also considered deviant behaviours as behaviours that depart from expectations. Those four views provided a useful framing of positive deviance that was used in varied disciplines and problem domains to understand the nature of deviations from the norm. This research employs both the statistical and normative view in defining PD. The statistical view would represent behaviours as a normal distribution where deviance lies at either extremes of the bell curve and the normative view will decide which extreme (right or left) reflects a positive behaviour.

Beyond these framings, the concept of PD evolved as a practical strategy in a number of domains, most notably in public health and international development. In 1976, the term PD was used for the first time to describe a practical strategy for the design of food supplementation programmes that are derived indigenously rather than extraneously through identifying dietary practises developed by mothers, in low-income families, having well-nourished children (S. M. Wishik & Van Der Vynckt 1976). The results of this study were not published, limiting uptake. It was not until the 1990s that positive deviance was introduced as a credible strategy for

academic and operational research in nutrition based on extensive observations and a strong emphasis on impact (Zeitlin 1991; Sternin et al. 1997; Sternin et al. 1998). The 1990s also witnessed its first large scale adoption in international development by Save the Children Foundation (SCF), which used PD as a strategy to reduce malnutrition in Vietnam, rehabilitating an estimated 50,000 malnourished children in 250 communities (Sternin 2002). But it has only really started to attract attention in the 2000s, when Sternin et al. introduced PD as an asset-based approach for social change and demonstrated how it can be operationalized as a domain-agnostic approach (Sternin & Choo 2000; Sternin 2002). Since then, PD has been applied across multiple domains, with public health being the most prominent.

Positive deviance can be understood as one thread in the strand of international development approaches that has sought to identify and disseminate "best practices"; i.e. solutions to development challenges that are seen to have worked well in at least one location and which are then sought to be replicated in other locations (e.g. Khennas & Barnett 2000; Williamson 2000; Oyen 2002). This has, for example, been the motivation behind large-scale development interventions such as the work of the Bill and Melinda Gates Foundation, and Jeffrey Sachs' Millennium Villages programme (Sachs 2008, Fejerskov 2017). Like the need-based approach mentioned above, this form of best practice approach has been highly-successful in rolling out technical solutions to development problems; for example, mass vaccination programmes to prevent measles in under fives, or bed-nets to reduce the incidence of malaria (Nanyunja et al. 2003; Agusto et al. 2013).

Where this approach has been less successful is in dealing with multi-faceted problems that incorporate not just the technical but also the social, the cultural and the political. For such problems, there have been two key concerns. First, that the best practices approach has been too "one size fits alls": identifying one solution and seeking to scale it out on a broad basis. This has led to problems of fit: failures of best practices because what worked well in one context did not fit into different contexts (Andrews 2012; Ramalingam et al. 2014). While still itself focused on "best practices", the positive deviance approach has been part of the response to these issues, seeking to derive solutions from, and scaling those solutions within, relatively bounded contexts which will enable a consistency of fit of those solutions, rather than seeking to import an external "best practice" (Singhal 2014). There have been instances of the traditional best

practices approach seeking to be more "bottom-up". For instance, Sachs's seminal five development interventions to end poverty were derived from a thorough investigation of village conditions and through learning and problem-solving along with the community members. However, the scale-out of these interventions has required significant and ongoing inputs of external development aid (Sachs 2006). The positive deviance approach has also been a response to this concern - by relying on locally-sourced solutions, implementation of positive deviance also tends to be relatively modest in terms of its resourcing requirements and more sustainable.

1.2 The Positive Deviance Approach

The PD approach, developed by Sternin (2002), can be defined as an asset-based approach to identify and promote uncommon behaviours and strategies employed by individuals, communities or organizations: "positive deviants" that outperform peers in solving intractable problems, despite sharing the same resource base. In what follows, the community will be the focal scope for discussion. The central premise of this approach is that it harnesses the inherent wisdom existing within a community to provide context-aware, affordable, efficient and sustainable solutions to pervasive problems. The five steps of the PD methodology can be outlined as follows (Positive Deviance Initiative 2010):

1. Defining the problem and determining desirable outcomes
2. Determining positive deviants i.e. individuals or communities who unexpectedly achieved the desired outcomes
3. Discovering the underlying practises and behaviours that led to those outcomes. This is referred to as Positive Deviance Inquiry. The identified behaviours should be accessible or economically feasible, unique i.e. different from the norm, and transferable.
4. Designing interventions to enable others to access and practice new behaviours
5. Monitoring and evaluating the PD intervention

There are a number of characteristics of the PD approach that have been identified in the literature. Positive deviants are unconsciously competent; that is, they usually do not know that they are different and that their practises are producing different results than their peers: they

simply "do not know what they know" (Pascale, Sternin & Sternin 2010). The PD approach works with problems that are embedded in the social fabric of communities, problems that require the disruption of social structures, cultural norms, or behaviours. It does not work with technical problems that can be solved by existing expert know-how, like developing a new medicine or building a bridge (Pascale et al. 2010; Felt 2011). Instead, it requires expertise in facilitating, discovering and mobilizing solutions which requires new know-how (Saïd Business School 2010). A third characteristic is that the desired outcome measure – i.e. the developmental change to be brought about – must be valid, concrete, widely endorsed and based on an accessible performance measure with adjustable covariates among individuals or communities. Additionally, it is important to have a substantial natural variation in performance between the positive deviants and the non-positive deviants (Bradley et al. 2009; Klaiman 2011), and to choose measures that are capable of indicating change in a timely manner (Felt & Cody 2011)

The mere existence of positive deviants is not enough; social processes and innovation dissemination mechanisms are crucial in promoting those uncommon behaviours among communities. Hence, the community is usually involved in the process of defining the problem, identifying positive deviants and practising the discovered behaviours instead of waiting for solutions to come from elsewhere (Bradley et al. 2009; Saïd Business School 2010; Pascale et al. 2010). This local engagement creates what can be referred to as self-efficacy (individuals' belief in their capacity to execute behaviours necessary to achieve a desired objective) which has been considered in many behavioural change models as a primary influencer in the adoption of recommended behaviours (Babalola et al. 2002; Babalola 2007). Finally, solutions are often localized: those that work in one place are not easily adoptable in another place for two reasons: 1) the inferred practises are particular to the circumstances of the intervention community so it might not be relevant to the circumstances and context of other communities, 2) its adoption depends largely on community self-efficacy, which is strengthened by the community's belief that solutions were derived from their own wisdom (Pascale et al. 2010; Singhal et al. 2010; Saïd Business School 2010).

According to the "Positive Deviance Initiative" (positivedeviance.org), directed by Monique Sternin, the successful application of PD has been reported in more than 60 countries across the globe with a total outreach of more than 30 million individuals for the period between 1990

and 2016. Applications include: reducing childhood malnutrition, enhancing school retention, eliminating neonatal mortality, limiting HIV transmission, improving salesforce productivity, fighting against female genital cutting (FGC), enhancing healthcare services, reducing transmission of antibiotic resistant bacteria (MRSA) in hospitals, and enhancing pregnancy outcomes. Here are a few success stories reported on the use of the PD approach:

- Reducing transmission of antibiotic resistant bacteria (MRSA) in three U.S. hospitals by 30-62% (Singhal et al. 2009; Pascale et al. 2010)
- Increasing student retention in primary schools in Misiones province, Argentina, by 50% (Dura & Singhal 2009; Pascale et al. 2010)
- Reducing girl trafficking in poor villages in East Java by 30% (Singhal & Dura 2009; Pascale et al. 2010)
- Reducing FGC in Egypt by 4% in three years (Pascale, Sternin & Sternin 2010)

International development agencies often seek to implement sustainable solutions for problems in developing countries by building local capacities. One measure of success can be the ability to implement projects that are capable of carrying on indefinitely without further support, be it financial or technical. This fits with the PD approach, which provides already proven solutions to problems based on local resources and know-how and generates local involvement in solving those problems (Levinson et al. 2007). Hence, it can reduce the risk of cultural resistance, the cost of developing and financing solutions and the reliance on external technical expertise. PD has been adopted widely by international development agencies like SCF, which incorporated PD as a cornerstone in its two-year poverty alleviation and nutrition program in Vietnam, reducing malnutrition by 65-80% (Sternin et al. 1997; Pascale et al. 2010). Other organizations that adopted the PD approach include, but are not limited to the Canadian International Development Agency (Ndiaye et al. 2009), United States Agency for International Development (Lapping and Schroeder 2002; Babalola et al. 2002; Marsh et al. 2002; Hendrickson et al. 2002; Kim et al. 2008), International Development Research Centre (Pant and Hambly Odame 2009; Roche et al. 2017), United Nations High Commissioner for Refugees (Lackovich-Van Gorp 2017) and the United Nations Children's Fund (Ahrari et al. 2002). This adoption came from their realization that positive deviants act as agents of change who can bridge the gap between expert knowledge systems and local knowledge systems (Pant & Hambly Odame 2009) and from their

belief that local expertise and indigenous wisdom is the best way to find culturally appropriate solutions to community problems.

1.2.1 Benefits of the PD approach

Some of the benefits of the PD approach that were identified from the reviewed literature include context awareness: that the PD approach takes into account local contextual variables providing solutions that are sensitive to local culture and relations and hence less vulnerable to community rejection (Wishik & Van Der Vynckt, 1976; Bradley et al. 2009; Pascale et al.; Singhal et al. 2010; Saïd Business School 2010; Klaiman 2011; Leavy 2011). The PD approach identifies affordable solutions, from the people, without requiring additional resources for their implementation (Pascale et al. 2010; Saïd Business School 2010). A third benefit is sustainability, as the PD approach provides already tested solutions derived from the people. This increases the likelihood of sustainability, since it makes solutions independent of development aid, external experts or donations (Pascale et al. 2010; Singhal et al. 2010; Lapping et al. 2002).

The PD approach facilitates adoption through practise. At-risk community members learn about successful behaviours through testing them not just seeing them, as it is easier to "act your way into a new way of thinking than to think your way into a new way of acting" (Pascale et al. 2010; Saïd Business School 2010; Singhal et al. 2010). Finally, a by-product of applying the PD approach is breaking social silos and creating social networks that did not exist before, through peer learning. PD fosters a sense of community by inviting and engaging members to achieve a certain goal. Hence, it promotes inclusion and cohesion. It can also create stronger relations between community members and various stakeholders e.g. government officials and NGOs (Saïd Business School 2010; Felt 2011).

1.2.2 Challenges and Research Gaps

Notwithstanding its potential benefits, the PD approach faces a number of challenges (hence, creating research opportunities that seek to address those challenges) that were identified based on a systematic literature review of empirical PD studies (Albanna & Heeks 2019) and a review of theoretical studies. The identified challenges can be summarised as follows:

- **Time:** The application of the PD approach is time consuming and sometimes the quality of implementation is compromised due to time constraints (Lapping et al. 2002; Marsh et al., 2004; Felt 2011; Albanna & Heeks, 2019); it takes months to complete the phases sequentially.
- **Identifying positive deviants:** Positive deviants are rare and costly to find with a typical prevalence rate of 0-10%; thus, being harder to find in smaller sample sizes (Marsh et al. 2004). And PD is generally applied on small-scale samples, because it relies mainly on primary data collection and observational methods to identify PD cases and those methods are time- and labour-intensive, with costs proportional to the size of the selected sample. Additionally, the small sample size limits the ability to identify deviance at different levels of aggregation e.g. community level deviance instead of individual level deviance (Albanna & Heeks 2019).
- **Static deviance:** Primary data collection provides a static picture of the population and the deviance within it since it depicts the behaviours of the analysed units at a certain point of time: deviance is therefore what is referred to as a point anomaly in statistics (Goldstein & Uchida 2016). Thus, the traditional approach lacks the ability to identify dynamic deviance represented in contextual anomalies where the abnormality of a data point is context specific i.e. it could be abnormal at some point of time and totally normal at another point of time; and in collective anomalies where abnormality is represented as a set of many data points (Albanna & Heeks 2019).
- **Behaviour identification risk:** The PD approach not only presumes that positive deviants' behaviours are invisible to neighbours, but also it presumes the willingness of positive deviants to share their strategies and best practices. But what if sharing compromises a certain competitive advantage (Felt & Cody 2011)? This also applies to organizations and their openness to share best practices (Bradley et al. 2009). Additionally, some of the studies that used observational methods in PD inquiry reported the Hawthorne effect (McCarney et al. 2007), which is the alteration of the behaviours of the subjects of a study due to being observed (Albanna & Heeks 2019).
- **Cost efficacy:** The cost-efficacy of the PD approach is sometimes limited due to the presumption that solutions identified at one place cannot be generalized across

populations (Marsh et al., 2004). It is therefore assumed that PD initiatives must be undertaken anew in each new context.

- **Resistance:** Traditional PD has sometimes seen a "natural human immune system rejection", a phrase coined by the Sternins to describe the behaviour of individuals who resist the adoption of new innovative practices, as old practices die hard, even if there is an obvious advantage from the new ones (Singhal et al. 2009)
- **Monitoring and evaluation:** The traditional PD approach lacks guidance and emphasis on how to apply credible monitoring and evaluation techniques (Schulz et al. 2010; Felt & Cody 2011). Evaluation may also fall into the selection bias trap, which is not knowing whether the PD approach actually worked due to the adopted strategies and practices, or because of other invisible associated factors that happened concurrently i.e. this is the problem of attributing causality to the PD intervention (Marsh et al. 2004; Karlan 2010).
- **Scaling:** PD relies heavily on community mobilization for the dissemination of successful behaviours and practises. It is time- and labour-intensive, with proven success in small scale adoption, but it makes large scale adoption a challenging task (Lapping et al. 2002; Marsh et al. 2004; Pascale et al. 2010; Felt 2011). However, there is a "need to test the assumption that positive deviance is, of necessity, a small scale approach by evaluating the effectiveness of different intensities of inquiry (number per population size)" (Marsh et al. 2004). There are also research questions that are not addressed yet, like: What are the quality issues that accompany scaling or replicating PD? Is there a critical mass needed for uptake? What are the key elements needed for scaling? And what is the cost of its wide-scale adoption on entire regions or countries? (Lapping et al. 2002; Karlan 2010; Saïd Business School 2010).
- **Narrow domain/geographic scope:** There is a clear gap in the domain and geographic adoption of PD applications in developing countries. Regarding the domain coverage, the vast majority – 89% - of reviewed PD empirical studies in developing countries were in public health, with 41% focused specifically on malnutrition. And to expand the use of the PD approach into new domains and disciplines, there is a need to develop context-specific frameworks to operationalize the process of adoption (Singhal et al. 2010; Herington and van de Fliert 2017). As for geographic coverage, there are nearly 150

developing countries (OECD 2017) but PD studies have encompassed only 20, with just four countries (India, Brazil, Pakistan, and Ethiopia) responsible for almost 50% of studies (Albanna & Heeks 2019).

Overall, we can see that the PD approach could be considered a powerful tool for international development that has been applied in a few fields to address complex problems requiring behavioural change. It offers a variety of potential benefits like context awareness, solution affordability and sustainability. However, it is faced with challenges that limit its uptake. Drawing out from the list above, one of the major challenges of the PD approach is the reliance on primary data collection to identify positive deviants and to infer their uncommon practices and behaviours. This makes it a time-and labour-intensive approach that is difficult to apply at large scale; despite the potential benefits that could result from large scale sampling, such as the ability to generalize practises inferred from larger samples of positive deviants and the ability to identify different types and aggregation levels of deviance. In this research we will explore how big data sources can be used in the PD approach to potentially mitigate this particular challenge and to explore other possible opportunities from their combined use.

1.3 Big Data for Development

The term "big data" was coined to describe the growing proliferation of data and our increasing ability to make productive use of it (Desouza & Smith 2014). Processing large scale data goes back to as early as 1890, where 15 million individual records were processed by the US census to improve governance, yet this is not considered big data according to the recent definition (Hilbert 2016). The primary difference is that today's data is produced at very high volume, velocity and variety, requiring new processing capacities, that may untap valuable knowledge. It was estimated that every person on Earth would create 1.7 MB of data every second in 2020, generating over 2.5 quintillion bytes of data daily (Domo 2018). Another important difference is that technological advancement converted data from analogue to digital, producing huge amounts of digital traces that can be analysed using advanced computation capabilities. During the last two decades, the world's capacity to exchange information grew from 0.3 exabytes (20% digitized) in 1986 to 65 exabytes (99.9% digitized) in 2007 (Hilbert & Lopez 2011). Similarly, the technological memory almost doubled every three years growing from 2.5 exabytes in 1986 to

300 exabytes in 2007 (94% digitized) (Hilbert & Lopez 2011). New technologies like wireless sensors, mobile devices, social network platforms and satellite imagery featured this technological advancement and have penetrated far across the world. For example, about 70% of the poorest fifth of the population in developing countries own a mobile phone (World Bank 2016), where the device's usage generates not just communication and location and mobility data but increasingly data related to health and livelihoods. In 2020, 59% of the world's population had access to the internet, with 4.57 billion active users. Of those people, 4.2 billion were active on mobile and 3.81 billion used social media (Domo 2020). This evolution and diffusion of digital infrastructures has led to the proliferation of data; and our increasing ability to make use of it, characterized in the enhanced processing and storage capacities that provide an opportunity to convert this data into knowledge that informs decisions.

Big data for development (BD4D) refers to the use of big data sources that are relevant to the policy and planning of development programmes. Its applications span a variety of domains and leverage new sources of data and new analytical tools. It differs from "traditional" development data in the form of surveys, and has been argued to be different from private sector "big data" that is defined mainly by its huge volume (United Nations Global Pulse, 2013). However, since this field is still in its operational infancy, a common definition for big data among the different BD4D practitioners and academics has been lacking. Having said this, one widely used definition was developed by the United Nations Global Pulse, an initiative which aims at accelerating the discovery, development and scaled adoption of big data innovation for sustainable development and humanitarian action. According to their definition, BD4D sources generally share some or all of the following features: 1) digitally generated i.e. data is machine readable, 2) passively produced i.e. it is a by-product of interactions with digital devices, 3) automatically collected i.e. there is a system that extracts and stores data as it is being generated, 4) geographically or temporarily trackable i.e. data has a spatial or temporal reference, and 5) continuously analysed i.e. data can be analysed in real-time (Pulse 2012).

If we follow the UN Global Pulse approach, then the term big data, in the context of development, might be misleading as size is not the defining feature. The difference to traditional datasets is mainly in the kind of data and how it is generated, captured and analysed.

1.3.1 Benefits of BD4D

Big data can fundamentally shift the way we pursue social change as it is capable of providing snapshots of the well-being of populations at high frequency, high degree of granularity, and from a wide range of angles, narrowing both time and knowledge gaps (Pulse 2012). In more detail, it therefore offers a number of potential benefits to international development. A key benefit is low cost of data with digital traces produced from digital platforms providing a low-cost alternative to traditional sources of data (survey data) by replacing variables of interests with correlated proxies (Hilbert 2016). This is evident, for example, in applications that use satellite images (Jean et al. 2016) and mobile call detail records (Blumenstock, Cadamuro & On 2015) to estimate the social and economic well-being of populations. Through monitoring populations, big data also makes it possible to understand where policy and programme interventions are failing, in order to make the necessary adjustments in a timely manner. In BD4D, "real time" does not necessarily mean occurring immediately; it means happening in a relatively short and relevant time permitting action to be taken in response (Pulse 2012).

Big data's geographical coverage can provide alternatives to random sampling. With a global penetration of 91% in the developed world and 90% in developing countries (Deloitte 2017), mobile phones can in theory sample much closer to the universe N instead of sampling n of the universe N (Hilbert 2016). Behavioural, socio-economic and demographic traits can be inferred from mobile phone records. For instance, call duration and frequency have been used to predict the level of socio-economic development in geographic regions (Frias-Martinez & Virseda 2013). Additionally, the ability of using and merging different sources of data reflecting a certain event or reflecting the behaviours of an individual, community or an organization provides a real-time, cross validated, fine grained picture of reality. An example is Thomson Reuters Market Psych Indices that extracts data from over three million articles and four million social media sites, covering 119 countries, every day to provide sentiment indices. Those indices provide much more detailed well-being and happiness indicators in comparison to the indicators used by the United Nations Human Development Index, like life expectancy, adult literacy, etc. (Hilbert, 2016; Pulse, 2012).

Finally, with the existence of big data, new processing and storage needs were mandated, and advanced analytics techniques came into existence. The real value of big data lies in our ability to use those techniques for better decision making. Big data made it possible to use techniques that perform much better if applied on huge amounts of data. An example is machine learning, a subfield of artificial intelligence, which gives computers the ability to learn from data without being explicitly programmed. Machine learning was often belittled during the 1990s, but has proven its significance during the 2010s by being applied to large amounts of data (Hilbert 2016). For example, big data analytics made it possible to identify 45 search terms that could predict flu outbreaks by processing around 450 million different mathematical models (Ginsberg et al. 2009). This would not have been possible using traditional models applied on small samples of data.

While these benefits do not exactly map to the challenges identified for traditional positive deviance – and while, as discussed in Chapter 2, big data comes with its own challenges – there are overlaps visible, suggesting some potential match of problem and solution, as discussed next.

1.4 Big Data and Positive Deviance

Identifying positive deviants is the essence of the PD approach; it is their existence that sparked the PD concept. However, as shown in section 1.2.2, positive deviants are rare to find with a low prevalence rate (0-10%) making their identification harder in small samples. Yet, PD is applied generally on small-scale samples because it relies on primary data collection for PD identification, and relies on community mobilization for practice dissemination, making the cost and time of any intervention a factor of the sample size. Those constraints add to the difficulty of identifying positive deviants, which would have been better suited to large-scale samples.

Drawing on the discussion in Section 1.3, we can see that big data provides an opportunity to potentially overcome this challenge in three main ways. In relation to size, big data could be used as a low-cost, large-scale alternative to small-scale primary data in the PD approach; increasing the odds of outlying deviants existing within the dataset, and reducing the time and

cost needed for identification. In terms of spatial and temporal resolution, the inherent characteristics of big data provide an opportunity to identify new types of anomalies (i.e. positive deviants) that were not identifiable using traditional primary data. For instance, big data could enable the identification of contextual anomalies and collective anomalies if, respectively, it was a near real time source of data covering long periods of time, and if it covered large geographic regions with different aggregation levels (individuals, communities, regions and countries). The third potential of big data arises from advanced data analytics. Machine learning-based approaches for anomaly detection outperform simple statistical models in identifying contextual and collective anomalies (Chandola et al. 2009; Goldstein and Uchida 2016). And since machine learning works better with larger datasets, big data would result in powerful machine learning models that could enhance anomaly detection. Machine learning could also be used to segment the PD intervention population (based on machine-inferred similarities) to target only the segments with socio-economic determinants similar to those of the positive deviants with segment-specific interventions. This would reduce both the time and cost required for scale-up.

Arising from these three identified potentials, in this research we propose to develop, investigate and apply a big data-based PD method which provides a two-sided opportunity. On the one side, this provides a new opportunity for positive deviance: using big data in the PD approach to identify positive deviants in new ways, dimensions and aggregation levels (although of course this would be limited to anomalies that can be digitally observed, mediated or recorded). On the other side, this provides a new opportunity to extract value from big data; applying the PD framework to big data to put knowledge about anomalies into action, as there is a "need to develop methodologies to characterize and detect socio-economic anomalies in context" (Pulse 2012). Additionally, big data-based PD could provide an analytical framework that enables the transformation of digital information into knowledge that informs better decisions. Previous literature in this field has noted that the biggest challenge of the "big data paradigm" is the ability to analyse this data for intelligent decision making (Hilbert 2013). One of the constraints on greater use of big data in PD (and of PD in BD4D) is the lack of a clear framework and methodology to guide their combined use, and this study attempts to fill this gap by developing, applying and validating a big data-based PD method.

1.5 Research Design

It is clear from the previous sections that PD is an effective problem solving approach for development, which is still facing challenges limiting its uptake. It also offers a possible framework for socio-economic anomaly detection that BD4D applications could benefit from. On the other hand, we can also see that big data, owing to its inherent characteristics, could offer an opportunity to enhance the current PD approach, provided there was an adapted conceptual PD framework that could guide its use. Hence, this research will test the combined use of big data and PD by developing a big data-based PD method and then applying it in practice.

1.5.1 Research Questions

In the previous sections I outlined the benefits and limitations of the PD approach and the use of big data for development. I also presented the opportunity that might arise from mixing both approaches. However, the value, the limitations and the means of combining them together are still unknown and this is what this thesis aims to answer. Hence, the research questions used to structure this empirical study were:

- How can we use big data in the positive deviance approach?
- What value will result from the use of big data in the positive deviance approach?
- What are the limitations of using big data in the positive deviance approach?

1.5.2 Research Aim and Objectives

This research aims to build on the established PD approach developed by Sternin (2002) to develop a big data-based PD method that will guide the use of big data and machine learning throughout the core stages of the PD methodology previously presented in section 1.2. The research objectives outline the specific steps needed to be taken to achieve the research aim and answer the research questions. To determine the means of using big data in the PD approach, there is a need to develop a method that accommodates the characteristics of big data and to validate this method through its application using different types of big data across different

domains and contexts. Additionally, and as noted in section 1.3.1, a key potential of big data is "big data analytics", characterised by the routine use of machine learning techniques. As such, this study aims to investigate this opportunity as well in relation to positive deviance. While there had been an aspiration to test out the new big data-based approach in all five stages of the PD approach, this ultimately was not possible because the last two (related to the design of interventions to disseminate PD practices and strategies, and the monitoring and evaluation of the effects of those intervention) would have required substantial time and resources that were well beyond the capacity of this PhD. Accordingly, the specific objectives focused on investigating the value of big data in relation to the first three stages; specifically, in the identification of positive deviants and in uncovering predictors of PD performance. Finally, and notwithstanding the positive potential identified from the literature, and as outlined above, there was also a need to investigate any limitations of the proposed method in order to ensure a critical evaluation of the its viability. The research objectives can thus be summarized as follows:

1. Develop and validate the proposed big data-based PD method
2. Identify machine learning techniques that can be utilized in the big data-based PD method
3. Determine the value of big data in the identification of positive deviants
4. Determine the value of big data in uncovering predictors of PD performance
5. Identify the limitations of the proposed big data-based PD method

1.5.3 Methodological Strategy

In 2015, I was undertaking a course on design thinking while working as a social business incubation officer. This course introduced me to the positive deviance approach, which I found extremely intriguing at that time. One year later, in 2016, I was developing a research proposal to apply for a PhD abroad. I wanted my research to lie at the intersection of data science (my second MSc degree), technology management (my first MSc degree) and social development (my working experience). I was drawn to the field of big data for development which achieves this multidisciplinarity but felt that there has been a bias towards substitution instead of complementarity when it comes to its application. I believed that big data will not be valuable

unless closely coupled with the human contribution that leads the process of transforming data into actionable insights. So instead of asking: what problems can be solved with big data? I asked the question: how can big data help us solve problems better? The starting point for me was to look for already existing development approaches that lend themselves to traditional data and complement this data with big data. At this point, I could not help but remember the positive deviance approach, about which I had learned in the previous year , and the analogy between the positive deviants of this approach and what we refer to as outliers in data. It occurred to me that perhaps big data might bring some new opportunities that could expand the use of positive deviance in development. This was the origin point for my whole study and a reflection of my overarching research philosophy: the philosophy of pragmatism that has guided the methodological choices for this study.

I am starting with a problem - how to integrate big data into the positive deviance approach - and I am aiming to contribute a practical solution that could inform future practice. As a pragmatist I will be working with different types of knowledge and methods that enable credible, well-founded, reliable and relevant data that advance the research (Kelemen & Rumens 2011). Within the pragmatist approach, mixed methods were my methodological choice in testing and iteratively developing the big data-based PD method presented in this thesis. Mixed methods are the branch of multiple methods research that combines the use of quantitative and qualitative data collection techniques and analytical procedures (Saunders, Lewis & Thornhill 2016). As a generalisation, quantitative data collection and analysis was used to identify deviance, and to test hypotheses about the bases of positive deviance generated from the qualitative data collection. This choice was also influenced by my pragmatic philosophical position, given that it is a widely-used methodology within pragmatist research (Feilzer 2010).

A multiple case study research strategy was chosen. Case study is a research strategy which investigates a particular phenomenon, empirically, using multiple methods and sources of evidence (Saunders, Lewis & Thornhill 2016). The rationale for selecting multiple cases is to check if findings are replicated across the different studies and at the same time test the method using different types of big data across different contexts. Hence, multiple case studies would support the generalization of findings and provide a richer, more complete assessment of the big data-based PD method. The logic governing the selection of the cases should reflect some

foundational interest of the research (Yin 2014). In this study, the systematic literature review (SLR), presented in Chapter Two, shaped my foundational interest, which was to cover the large range of variability in the applications of BD4D in terms of the tracked elements, big data sources, units of analysis and their domain and geographic distribution as highlighted below.

1. **Tracked elements:** In the SLR, BD4D studies were classified into four main groups based on the elements being tracked (locations, words, nature and economic activity) according to the taxonomy of big data applications proposed by Hilbert (2016). The case studies should ensure a diversity of the elements being tracked rather than being restricted to just one type.
2. **Big data sources:** The SLR revealed that remote sensing data, mobile data and online data are the most common data sources in BD4D applications. The case studies should thus ensure a diversity of big data sources used. In practice, it proved very challenging to gain access to mobile data but case study sources were able to cover both online and remote sensing data.
3. **Unit of analysis:** In the SLR, BD4D applications covered different aggregation levels starting with individuals and then communities up to broader geographical units such as areas or districts. The case studies selected should thus ensure some diversity in the aggregation level being investigated.
4. **Domain distribution:** The SLR presented the diversity of the domains covered by the BD4D applications ranging from economics, public health and environmental studies to poverty measurement and infectious disease control. The case studies should ensure a diversity of application domains – ideally branching beyond the existing PD concentration of public health – in order to demonstrate the ability of big data to expand the scope of PD.
5. **Geographic distribution:** The investigated PD phenomenon should be in a developing country context, since the focus of this study is on the application of the PD approach in development, but the case studies should ideally be drawn from developing country locations other than those four (India, Brazil, Pakistan, and Ethiopia) which have dominated existing PD studies.

Given the origin point for my study came to me while sitting in my office in Cairo, I wanted my first application of this new method to be an Egyptian case study, if possible. This was also logical given my understanding of the Egyptian context, and my wealth of local contacts. As

such, the method was first tested in a case study of Egyptian researchers who outperformed their peers in terms of research outputs. Following that, the remaining cases were conducted as part of a global initiative that I collaboratively created with the GIZ Data Lab, UN Global Pulse Lab Jakarta (PLJ) and the UNDP Accelerator Labs Network (Data Powered Positive Deviance 2020). This initiative emerged after the publication of the literature review and introduction of the idea of big data-based positive deviance that forms Chapter Two of this thesis (Albanna & Heeks 2019). Catherine Vogel (the GIZ Data Lab Coordinator) came across the paper which captured her interest and encouraged her to raise funds to scale the methodology through multiple projects with partners from development and academia. She reached out to myself and my supervisor, Richard Heeks and we supported her in the co-development of a funding proposal to GIZ which led to 750,000 euros being raised directly for the implementation of the projects, with an equivalent sum being contributed by the UNDP Accelerator Labs: a €1.5m project in total.

The research strategy for each individual project or case study was action research. Action research promotes organizational learning to produce practical outcomes through the identification of issues, planning of action, taking of action, and evaluation of action (Saunders, Lewis & Thornhill 2016). It is seen as particularly appropriate for research guided by pragmatism (Reason 2003) because it bridges the gap between research and practice by integrating, rather than chronologically separating, the two processes of research and action (Somekh 1995). It would therefore allow the application of the big data-based PD method to be fed back into its conceptualisation; that re-conceptualisation then refining practice in an iterative cycle. Our aim was to run a number of pilot projects to check if and how the big data-based PD method can be used to tackle development challenges.

Five action research projects were conducted as part of this initiative by the different partners in five developing countries to identify and understand: farmers achieving higher than usual cereal crop productivity in Niger and Indonesia; cattle farmers in Ecuador who are deforesting below average rates; public spaces in Mexico City where women are safest; and communities in Somalia which are able to preserve their rangelands despite the frequent droughts. I was the methodological, intellectual and analytical lead in those projects and also the key advisor. Since March 2020 to date, I have held weekly calls with all the projects to guide them through the

implementation of the method and to capture and transfer learnings as we undergo the different stages of the big data-based PD method. This collaboration enabled me to test the method with real life development problems and practitioners in a much wider domain space and with a focus on method development and refinement, instead of being solely involved in post-hoc analysis of case study evidence. Such a broad role required extensive resources and time that are beyond the means of an independent research investigator.

Table 1 shows how the aspects of the aforementioned criteria were covered in my multiple-case study research strategy. The cases offered diversity in terms of big data types: remote sensing and online data. This was supported by a variety of tracking elements: words, economic activity, nature and locations. The units of analysis covered different aggregation levels starting with individuals, farms and communities up to geographical units representing urban areas and villages. All of this took place in six different countries. This diversity of domains, countries, data and scales was seen as important in helping to broaden the testing base for the big data-based PD method and to strengthen its likely generalisability.

Case Study/Action Research Project	**Selection Criterion**				
	Tracked Elements	**Big data Source Used**	**Domain**	**Country**	**Unit of Analysis**
1. Researchers who outperformed their peers in terms of research outputs	Words	Online data	Research	Egypt	Individuals
2. Rice-growing villages achieving higher than usual crop productivity	Economic activity	Remote sensing data	Agriculture	Indonesia	Villages
3. Cattle farmers who are	Nature			Ecuador	Farms

deforesting below average rates		Remote sensing data	Cattle Ranching		
4. Public spaces with lower than expected violence against women	Location	Online data	Crime Control	Mexico	Urban Blocks
5. Pastoral communities that preserve healthy rangelands despite recurring droughts	Nature	Remote sensing data	Rangeland Management	Somalia	Community Rangelands
6. Drought resilient cereal growing villages with rain-fed agricultural systems	Economic activity	Remote sensing data	Climate Resilience	Niger	Communities

Table 1: Case study and selection criteria matrix

1.6 Structure and Content

I chose to present my thesis in journal format, since the nature of the PhD study allows the findings to be presented in a series of linked papers with suitable length and quality of publication. The empirical data collected also enabled distinctive methodological contributions and analytical findings. The thesis consists of three papers and one report, each appearing as a

separate chapter (Chapters Two, Three, Four and Five). I end with a discussion in Chapter Six which draws together the findings, with specific reference to each of the research questions.

The first paper "Positive deviance, big data, and development: A systematic literature review" (Albanna & Heeks 2019) was published in the peer-reviewed journal, The Electronic Journal of Information Systems in Developing Countries. It is presented as Chapter Two and analyses the current state of PD and the potential for big data to address some of the challenges facing it. In this paper I have undertaken a systematic literature review of the empirical applications of PD and big data in developing country contexts to answer three main questions: 1) how is PD currently being applied in development?; 2) how is big data currently being applied in development?; and 3) what development value might result from the combined use of big data and the PD approach?. The final corpus of analysis included 41 PD articles and 34 big data articles. This paper provides a thematic classification of the PD studies, with a particular focus on the challenges arising from work to date. Similarly, big data studies were classified based on their application in development, also adding a discussion on challenges that could limit its use in the PD approach. I then interrogate the challenges facing the PD approach while presenting how big data can respond them. At the end of the paper I conceptualize a preliminary framework that could guide the use of big data in the PD approach. This framework was first named "big data-based positive deviance" but I later on called it "data-powered positive deviance" (DPPD) due to its reliance on other forms of data that are both traditional and non-traditional, but are not necessarily "big". I was corresponding and first author for this publication, with my primary supervisor, Richard Heeks, as second author, who offered critique of my drafts in addition to reviewing and editing the paper content.

The second paper "Publication outperformance among global South researchers: An analysis of individual-level and publication-level predictors of positive deviance" (Albanna, Handl & Heeks 2021) was published in the peer-reviewed journal, Scientometrics. It constitutes Chapter Three of this thesis and is a case study on positive deviance in research performance where the DPPD method was applied and tested for the first time to identify and characterize publication outperformance in a Global South country. In this paper I ask: who are those outperformers or positive deviants, what are their characteristics and how were they able to overcome a few of the challenges facing Southern researchers? I examined a sample of 203 Information System

researchers in Egypt who were classified into positive deviants and non-positive deviants by analysing their publications and citations on Google Scholar. I was able to identify and cluster 26 positive deviants based on six citation metrics, and their underlying attributes, attitudes, behaviours, and publications were examined using a mixed methods approach. I started with an inductive inquiry to generate hypotheses from a small sample of positive deviants followed by a deductive inquiry to identify significant differences between positive deviants and non-positive deviants. I used a combination of data sources (interviews, surveys and publications) and analytical techniques (partial least square regression and topic modelling) to uncover the factors underlying the PD performance. Through this study I identified individual-level and publication-level predictors of positive deviance in a global South context that had not been identified in previous studies, and was able to provide pointers to ways of overcoming challenges specific to Southern researchers. I was corresponding and first author for this paper, with Julia Handl my second supervisor as second author, and Richard Heeks my first supervisor as third author. They offered critique of my drafts, reviewed and edited the paper content and provided guidance on the data collection and analysis methods.

The technical report "Identifying potential positive deviants across rice producing areas in Indonesia: An application of big data analytics and approaches" (Albanna, Dhar Burra & Dyer 2020) was published as a peer-reviewed technical report of the United Nations Pulse Lab Jakarta. It is presented as Chapter Four and summarises the DPPD research project I conducted during my fellowship at UN Pulse Lab Jakarta. This project is my second attempt to test and validate the DPPD method where we detected positive deviance in the agricultural domain by merging big earth observation (EO) data with official statistical and administrative data. The report presents a stepwise approach for the identification and validation of positively deviant rice villages in Indonesia i.e. villages with significantly higher agricultural productivity in comparison to neighbouring villages with similar socio-economic and environmental conditions. The study sample included 17,517 villages out of which we were able to identify 141 potential positive deviants using univariate and multivariate outlier detection techniques. Following that, the identified potential positive deviants were validated through visual inspection using Google Earth time scale tool, time series analysis using EO data at the pixel level, and bivariate analysis using official statistics. The report details the results, key learnings, along with some actionable recommendations for future work in the agriculture domain. I was

lead and corresponding author for this report. Dharani Dhar Burra, who was previously an agricultural and food systems data scientist at PLJ was second author, and Michael Dyer who was previously a geospatial information systems officer at PLJ was third author. I led the writing of this report, developed the main conceptual idea and did the majority of the data analysis. Dharani contributed to the development of the method and the selection of the analytical approaches used, which required his domain expertise. He also extracted all the EO data that was used in the analysis. Michael Dyer was the project manager and he forged necessary partnerships to obtain the official statistical and administrative data that was used in the analysis. Both Dharani and Michael undertook EO data analysis for the validation of the identified potential positive deviants. I conducted bivariate analysis using official statistics for the validation. A detailed authors' contributions paragraph is provided in the report.

The third paper "Data-powered positive deviance: Combining traditional and non-traditional data to identify and characterise development-related outperformers" was published in the peer-reviewed journal, Development Engineering. It constitutes Chapter Five, which presents the main contribution of this thesis. This paper provides a first exposition of the DPPD method, by describing a methodological framework that guides the combined use of traditional data sources and non-traditional digital data sources to identify and characterize positive deviants in development-related challenges. The DPPD method developed iteratively through its application in the case study presented in Chapter Three, followed by the PLJ project presented in Chapter Four, and the four additional GIZ-/UNDP-funded action research projects conducted in Ecuador, Mexico, Niger and Somalia outlined in Table 1. These projects were part of the DPPD initiative and I was the lead technical consultant. I co-authored this paper with Richard Heeks, Andreas Pawelke, Jeremy Boy, Julia Handl and Andreas Gluecker. The writing of this paper is mainly mine and I am the corresponding and first author. Richard Heeks and Julia Handl - my supervisors - offered critique of my drafts, reviewed and edited the paper content and provided guidance on the paper structure. Andreas Pawelke is a consultant for the GIZ Data Lab on the DPPD initiative, and he contributed to the 'Assessing Problem-Method Fit' section of the paper and reviewed the paper. Jeremy Boy was the project manager from the UNDP Accelerator Labs side and Andreas Gluecker was the project manager from the GIZ side. Both of them reviewed the final version of the paper and ensured the alignment of the findings with

the project outputs. A brief summary of each of those four additional GIZ/UNDP projects is outlined below:

- **Ecuador cattle farming project** (Grijalva et al. 2021): Deforestation is an alarming problem for the Amazon region, where 99% of the deforested areas are transformed into agricultural land, 64% of which is used as pastureland for livestock faming (Ministry of Envirnoment of Ecuador 2016). This project searches for cattle farmers in the Ecuadorian Amazon who operate in areas of potential forest clearance for farming without themselves contributing to deforestation. Positive deviants are defined as cattle-raising farms with deforestation rates that are significantly lower than what might be expected based on the size of the farm, the land use, soil adaptability and cattle density. For the identification of potential positive deviants, we used land cover and land use data derived from satellite imagery, as well as climate, soil, socio-economic and vaccination data. Regression analysis techniques (generalized linear models) were used to identify outliers or 'potential' positive deviants based on their residuals i.e. the difference between what is observed and what is predicted given their contextual conditions. The study sample covered two cantons in Ecuador, Joya de los Sachas and Sucúa, where 5,332 and 5,701 farms were inspected respectively. We were able to identify 53 potential positive deviants across both cantons who were targeted for fieldwork using qualitative methods to uncover their uncommon practices, attributes and attitudes. Our aim in this project is to scale those practices in order to reduce the negative effects of livestock farming on deforestation.
- **Mexico safe public spaces project** (Cervantes, Rios & Soto 2021): Violence against women and girls (VAWG) in Mexico is a pressing problem. Two thirds of all girls and women above the age of 15 have reported experiencing at least one incident of violence in their lifetime (National Institute of Statistics and Geography 2017). VAWG occurs in the street, parks and, to a lesser extent, on buses, minibuses and the subway. In this project, we used geospatial data, official statistics, administrative data and publicly available crime reports and 911 calls to identify positively deviant public spaces that have lower than expected incidence of outdoor VAWG when compared to spaces with similar traffic, socio-demographic and economic conditions. Multiple regression analysis techniques (linear regression, negative binomial and lasso regression) were also used to identify potential positive deviants. The

study sample covered 2,431 urban blocks in Mexico City, out of which 18 turned out to be potential positive deviants. Those PD urban blocks will be further investigated to explore factors contributing to lower rates of outdoor VAWG. Initial predictors of PD performance were identified using secondary data sources and discriminant analysis techniques. Our goal is to inform local policy makers and grassroots initiatives on those factors, thereby contributing to the design of safer public spaces in Mexico City.

- **Niger agricultural project** (Gluecker, Lehman & Barvels 2021): Agriculture in Niger is under tremendous pressure because of climate change, related reduction of rainfall affecting crop cycles, and local insecurity. For example, 1.7 million people are estimated to be food insecure in 2021 because of the challenges to agriculture (World Food Programme 2021). The project in Niger aims at identifying and scaling practices of positively deviant cereal-growing communities that produce higher than expected yields of sorghum and pearl millet despite the negative influences of droughts and conflict. We used readily available earth observation, geospatial and administrative data to identify those communities. Regression and machine learning techniques (linear regression, neural nets and XG boost) were used to identify potential positive deviants. The study sample covered 12,093 communities in the Sahelian region, where there is a predominance of rain-fed agriculture. We identified 180 communities that were performing particularly well, which we validated and are currently planning for the fieldwork stage which was preceded by a pilot investigation that uncovered some interesting practices and strategies. Our goal is to identify and leverage the local practices and strategies driving this outperformance to inform the design of interventions that can support communities in increasing their agricultural productivity despite climate change effects.
- **Somalia rangelands project** (Abdullahi, Albanna & Barvels 2021): Repeated droughts in Somalia between 2010 and 2017 have caused more than a quarter of a million deaths, and contributed to the displacement of roughly 4.2 million people (Care 2021). Pastoralists and their livelihoods suffer most from the burden of this climate change-exacerbated phenomenon. The project in Somalia aims to identify positively deviant pastoral communities that are able to sustain the health of their surrounding rangelands despite the recurring droughts. In this project we have used earth observation data along with administrative land cover data to identify community rangelands that are performing better

than rangelands under similar local conditions. To identify communities with positive vegetation development, we applied a method called local scaling. This method measures vegetation condition by comparing it with a reference value representing the state expected in the absence of extreme climate and human land use. The study sample covered 314 communities in the West Gollis area of Somaliland, which is a zone with a majority pastoral community. We identified a total of 13 communities as positive deviants and these will be explored further in fieldwork to uncover the factors underlying their outperformance. The aim of the project is to build on local knowledge to help communities understand how they might preserve their rangelands in order to maintain their pastoral livelihoods.

In Chapter Six, I draw together and discuss the overall findings and synthesise the response to my research questions. Finally, within that chapter, I set out my claim to be making a contribution to the body of knowledge and make some recommendations for data and development policy and practice, and future research directions.

References

Abdullahi, H., Albanna, B. & Barvels, E. (2021) *Rangelands Defying the Odds: A Data Powered Positive Deviance Inquiry in Somalia*, *Data Powered Positive Deviance*. Available at: https://dppd.medium.com/rangelands-defying-the-odds-a-data-powered-positive-deviance-inquiry-in-somalia-90772de392dd (Accessed: 4 August 2021).

Agusto, F. B., Del Valle, S. Y., Blayneh, K. W., Ngonghala, C. N., Goncalves, M. J., Li, N., ... & Gong, H. (2013). The impact of bed-net use on malaria prevalence. *Journal of Theoretical Biology*, *320*, 58-65.

Ahrari, M., Kuttab, A., Khamis, S., Farahat, A.A., Darmstadt, G.L., Marsh, D.R. & Levinson, F.J. (2002) Factors associated with successful pregnancy outcomes in upper Egypt: A positive deviance inquiry, *Food and Nutrition Bulletin*, 23(1), 83–88.

Albanna, B., Dhar Burra, D. & Dyer, M. (2020) *Identifying Potential Positive Deviants (PDs) Across Rice Producing Areas in Indonesia: An Application of Big Data Analytics and Approaches*. Jakarta: Pulse Lab Jakarta.

Albanna, B., Handl, J. & Heeks, R. (2021) Publication outperformance among global South researchers: An analysis of individual-level and publication-level predictors of positive deviance, *Scientometrics,* 1-57.

Albanna, B. & Heeks, R. (2019) Positive deviance, big data, and development: A systematic literature review, *The Electronic Journal of Information Systems in Developing Countries*, 85(1), e12063.

Andrews, M. (2012). The logical limits of best practice administrative solutions in developing countries. *Public Administration and Development*, *32*(2), 137-153.

Babalola, S. (2007) Motivation for late sexual debut in Cote d'Ivoire and Burkina Faso: A positive deviance inquiry, *Journal of HIV/AIDS Prevention in Children and Youth*, 7(2), 65–87.

Babalola, S., Awasum, D. & Quenum-Renaud, B. (2002) The correlates of safe sex practices among Rwandan youth: A positive deviance approach, *African Journal of AIDS Research*, 1(1), 11–21.

Blumenstock, J., Cadamuro, G. & On, R. (2015) Predicting poverty and wealth from mobile phone metadata, *Science*, 350(6264), 1073–1076.

Bradley, E. H., Curry, L. A., Ramanadhan, S., Rowe, L., Nembhard, I. M. & Krumholz, H. M. (2009) Research in action: Using positive deviance to improve quality of health care, *Implementation Science*, 4(1), 25.

Care (2021) *Somalia Food Insecurity Crisis, Care*. Available at: https://www.care.org/our-work/disaster-response/emergencies/somalia-food-insecurity-crisis/ (Accessed: 4 August 2021).

Cervantes, A., Rios, G. & Soto, I. (2021) *Identifying Safe(r) Public Spaces for Women in Mexico City*, *Data Powered Positive Deviance*. Available at: https://dppd.medium.com/identifying-safe-r-public-spaces-for-women-in-mexico-city-4f3d49d269d6 (Accessed: 4 August 2021).

Chandola, V., Banerjee, A. & Kumar, V. (2009) Anomaly detection: A survey, ACM *Computing Surveys*, 41(3), 1–58.

Data Powered Positive Deviance (2020) *Launching the Data Powered Positive Deviance Initiative, Data Powered Positive Deviance*. Available at: https://dppd.medium.com/ (Accessed: 4 August 2021).

Deloitte (2017) *Global Mobile Consumer Trends , 2nd edition*. London: Deloitte.

Desouza, K. & Smith, K. (2014) Big data for social innovation, *Stanford Social Innovation Review*, *12*(3), 38-43.

Dodge, D. L. (1985) The over-negativized conceptualization of deviance: A programmatic exploration, *Deviant Behavior*, 6(1), 17–37.

Domo (2018) *Data Never Sleeps 6.0, Domo*. Available at: https://www.domo.com/learn/infographic/data-never-sleeps-6 (Accessed: 4 August 2021).

Domo (2020) *Data Never Sleeps 8.0, Domo*. Available at: https://www.domo.com/learn/infographic/data-never-sleeps-8 (Accessed: 4 August 2021).

Dura, L. & Singhal, A. (2009) Will Ramon finish sixth grade? Positive deviance for student retention in rural Argentina, *Positive Deviance Wisdom Series*, 2, 1–8.

Ewald, K. & Jiobu, R. (1985) Explaining positive deviance: Becker's model and the case of runners and bodybuilders, *Sociology of Sport Journal*, 2, 144–156.

Fejerskov, A. M. (2017). The influence of established ideas in emerging development organisations: Gender equality and the Bill and Melinda Gates Foundation. *The Journal of Development Studies*, *53*(4), 584-599.

Felt, L. J. & Cody, M. (2011) *Present Promise, Future Potential: Positive Deviance and Complementary Theory,* Unpublished manuscript. Available at: http://www.laurelfelt.org/wp-content/uploads/2011/06/PositiveDeviance-CodyMayer.LaurelFelt.Quals_.May2011.pdf (Accessed: 21 September 2021).

Frias-Martinez, V. & Virseda, J. (2013) Cell phone analytics: Scaling human behavior studies into the millions, *Information Technologies & International Development*, 9(2), 35–50.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S. & Brilliant, L. (2009) Detecting influenza epidemics using search engine query data, *Nature*, 457(7232), 1012–1014.

Gluecker, A., Lehman, E. & Barvels, E. (2021) *Searching for Positive Deviants Among Cultivators of Rainfed Crops in Niger*, *Data Powered Positive Deviance*. Available at: https://dppd.medium.com/searching-for-positive-deviants-among-cultivators-of-rainfed-crops-in-niger-8dbbcceaf4ec (Accessed: 4 August 2021).

Goldstein, M. & Uchida, S. (2016) A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, *PLoS ONE*, 11(4), e0152173.

Grijalva, A., Jiménez, P., Albanna, B. & Boy, J. (2021) *Deforestation, Cows, and Data: Data Powered Positive Deviance Pilot in Ecuador's Amazon*, *Data Powered Positive Deviance*. Available at: https://dppd.medium.com/deforestation-cows-and-data-data-powered-positive-deviance-pilot-in-ecuador-s-amazon-648aa0de121c (Accessed: 3 August 2021).

Hendrickson, J. L., Dearden, K., Pachón, H., An, N. H., Schroeder, D. G. & Marsh, D. R. (2002) Empowerment in rural Viet Nam: Exploring changes in mothers and health volunteers in the context of an integrated nutrition project, *Food and Nutrition Bulletin*, 23(4), 86–94.

Herington, M. J. & van de Fliert, E. (2018) Positive deviance in theory and practice: A conceptual review, *Deviant Behavior*, 39(5), 664–678.

Hilbert, M. (2013) Big data for development: From information-to knowledge societies, *SSRN Electronic Journal*, 1–39.

Hilbert, M. (2016) Big data for development: A review of promises and challenges, *Development Policy Review*, 34(1), 135–174.

Hilbert, M. & Lopez, P. (2011) The world's technological capacity to store, communicate, and compute information, *Science*, 332(6025), 60–65.

Hughes, R. & Coakley, J. (1991) Positive deviance among athletes: The implications of overconformity to the sport ethic, *Sociology of Sport Journal*, 8(4), 307–325.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. & Ermon, S. (2016) Combining satellite imagery and machine learning to predict poverty, *Science*, 353(6301), 790–794.

Karlan, D. (2010) Survival of the deviant, *Stanford Social Innovation Review,* 4(8), 17-18.

Kelemen, M. & Rumens, N. (2011) *An Introduction to Critical Management Research.* London: Sage.

Khennas, S., & Barnett, A. (2000). *Best Practices for Sustainable Development of Micro Hydro Power in Developing Countries*, ESMAP Technical Paper 006. Washington, DC World Bank.

Kim, Y. M., Heerey, M. & Kols, A. (2008) Factors that enable nurse-patient communication in a family planning context: A positive deviance study, *International Journal of Nursing Studies*, 45(10), 1411–1421.

Klaiman, T. (2011) Learning from top performers using a positive deviance approach, *American Journal of Medical Quality*, 26(6), 422–422.

Lackovich-Van Gorp, A. (2017) Unearthing local forms of child protection: Positive deviance and abduction in Ethiopia, *Action Research*, 15(1), 39–52.

Lapping, K. & Schroeder, D. (2002) Comparison of a positive deviant inquiry with a case-control study to identify factors associated with nutritional status among Afghan refugee children in Pakistan, *Food & Nutrition,* 23(4), 26–33.

Leavy, B. (2011) Leading adaptive change by harnessing the power of positive deviance, *Strategy & Leadership*, 39(2), 18–27.

Levinson, F. J., Barney, J., Bassett, L. & Schultink, W. (2007) Utilization of positive deviance analysis in evaluating community-based nutrition programs: An application to the Dular program in Bihar, India, *Food and Nutrition Bulletin*, 28(3), 259–265.

Marsh, D.R., Sternin, M., Khadduri, R., Ihsan, T., Nazir, R., Bari, A. & Lapping, K. (2002) Identification of model newborn care practices through a positive deviance inquiry to guide behavior-change interventions in Haripur, Pakistan, *Food and Nutrition Bulletin*, 23(4), 109–18.

Marsh, D. R., Schroeder, D. G., Dearden, K. A., Sternin, J. & Sternin, M. (2004) The power of positive deviance, *British Medical Journal*, 329(7475), 1177–1179.

McCarney, R., Warner, J., Iliffe, S., Van Haselen, R., Griffin, M. & Fisher, P. (2007) The Hawthorne Effect: A randomised, controlled trial, *BMC Medical Research Methodology*, 7.

Ministry of Envirnoment of Ecuador (2016) *Forests for Good Living - Ecuador REDD + Action Plan (2016-2025)*. Quito: Governement of Ecuador - Ministry of Environement

Nanyunja, M., Lewis, R. F., Makumbi, I., Seruyange, R., Kabwongera, E., Mugyenyi, P., & Talisuna, A. (2003). Impact of mass measles campaigns among children less than 5 years old in Uganda. *Journal of Infectious Diseases*, *187*(Supplement 1), S63-S68.

National Institute of Statistics and Geography (2017) *National Survey on the Dynamics of Household Relations 2016*. Mexico City: National Institute of Statistics and Geography .

Ndiaye, M., Siekmans, K., Haddad, S. & Receveur, O. (2009) Impact of a positive deviance approach to improve the effectiveness of an iron-supplementation program to control nutritional anemia among rural Senegalese pregnant women, *Food and Nutrition Bulletin*, 30(2), 128–136.

Nel, H. (2018) A comparison between the asset-oriented and needs-based community development approaches in terms of systems changes, *Practice*, 30(1), 33–52.

Osborne, J. W. & Overbay, A. (2004) The power of outliers (and why researchers should always check for them), *Practical Assessment, Research & Evaluation*, 9(6), 1–8.

Øyen, E. (ed.) (2002). *Best Practices in Poverty Reduction: An Analytical Framework*. London: Zed Books.

Pant, L. P. & Hambly Odame, H. (2009) The promise of positive deviants: Bridging divides between scientific research and local practices in smallholder agriculture, *Knowledge Management for Development Journal*, 5(2), 160–172.

Pascale, R., Sternin, J. & Sternin, M. (2010) *The Power of Positive Deviance: How Unlikely Innovators Solve the World's Toughest Problems*. Boston, Massachusetts: Harvard Business Press.

Positive Deviance Initiative (2010) *Basic Field Guide to the Positive Deviance Approach*. Medford, MA: Positive Deviance Initiative, Tufts University.

Pulse, U.N.G (2012) *Big Data for Development: Challenges & Opportunities*. New York: UN Global Pulse.

Ramalingam, B., Laric, M., & Primrose, J. (2014). *From Best Practice to Best Fit: Understanding and Navigating Wicked Problems in International Development*. London: Overseas Development Institute, London.

Reason, P. (2003) Pragmatist philosophy and action research: Readings and conversation with Richard Rorty, *Action Research*, 1(1), 103–123.

Roche, M. L., Marquis, G. S., Gyorkos, T. W., Blouin, B., Sarsoza, J. & Kuhnlein, H. V. (2017) A community-based positive deviance/hearth infant and young child nutrition intervention in ecuador improved diet and reduced underweight, *Journal of Nutrition Education and Behavior*, 49(3), 196-203.

Sachs, J. (2006). *The end of poverty: Economic possibilities for our time*. London, UK: Penguin.

Sachs, J. (2008), The end of poverty: economic possibilities for our time, *European Journal of Dental Education*, 12, pp.17-21.

Saïd Business School (2010) *Exploring Positive Deviance – New Frontiers in Collaborative Change*. Oxford, UK: Saïd Business School, University of Oxford.

Saunders, M., Lewis, P. & Thornhill, A. (2016) *Research Methods for Business Students*. London, United Kingdom: Pearson Education.

Schulz, K. F., Altman, D. G. & Moher, D. (2010) CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials, *Trials*, 11(1), pp.1-8.

Singhal, A. (2014). *The positive deviance approach to designing and implementing health communication interventions. Global Health Communication Strategies in the 21st Century: Design, Implementation, and Evaluation*. New York, NY: Peter Lang Publishing Group, 174-89.

Singhal, A., Buscell, P. & McCandless, K. (2009) Saving lives by changing relationships: Positive deviance for MRSA prevention and control in a US hospital, *Positive Deviance Wisdom Series*, 3(3), 1–8.

Singhal, A. & Dura, L. (2009) *Protecting Children from exploitation and trafficking: Using positive deviance approach in Uganda and Indonesia*. Texas: Department of Communication, The University of Texas.

Singhal, A. (2011) Turning diffusion of innovation paradigm on its head: the positive deviance approach to social change, in *The Diffusion of Innovations*, A. Vishwanath & G. A. Barnett (eds). New York, NY: Peter Lang, 193–205.

Singhal, A., Sternin, J. & Dura, L. (2009) Combating malnutrition: Positive deviance grows roots in Vietnam in the land of a thousand rice fields, *Positive deviance wisdom series*, 1, 1–8.

Spreitzer, G. M. & Sonenshein, S. (2003) Positive deviance and extraordinary organizing, *Upward Spiral & Positive Change*, 207, 224.

Spreitzer, G. M. & Sonenshein, S. (2004) Toward the construct definition of positive deviance, *American Behavioral Scientist*, 47(6), 828–847.

Sternin, J. (2002) Positive deviance: A new paradigm for addressing today's problems today, *The Journal of Corporate Citizenship*, 57–63.

Sternin, J. & Choo, R. (2000) The power of positive deviancy, *Harvard Business Review*, 78(1), 1–3.

Sternin, M., Sternin, J. & Marsh, D. (1998) *Designing a Community-Based Nutrition Program Using the Hearth Model and the Positive Deviance Approach: A Field Guide.* Westport, CT: Save the Children.

Sternin, M., Sternin, J. & Marsh, D. (1997) *Rapid Sustained Childhood Malnutrition Alleviation Through a Positive-Deviance Approach in Rural Vietnam: Preliminary Findings*. Arlington, VA: Partnership for Child Health Care, BASICS.

Warren, D. E. (2003) Constructive and destructive deviance in organizations, *Academy of Management Review*, 28(4), 622–632.

Williamson, I. P. (2000). *Best Practices for Land Administration Systems in Developing Countries*. Washington, DC: World Bank.

Wishik, S. M. & Van Der Vynckt, S. (1976) The use of nutritional 'positive deviants' to identify approaches for modification of dietary practices, *American Journal of Public Health*, 66(1), 38–42.

Wolfgang, M. E. (1965) *Social Deviance: Social Policy, Action and Research*. New York: JSTOR.

World Bank Group (2016) *World Development Report 2016: Digital Dividends.* Washington, DC.: World Bank Publications.

World Food Programme (2021) *WFP Niger Country Brief.* Niger: World Food Programme.

Yin, R. K. (2014) *Case Study Research: Design and Methods (Fifth).* London, UK: SAGE Publications Ltd.

Yvonne Feilzer, M. (2010) Doing mixed methods research pragmatically: Implications for the rediscovery of pragmatism as a research paradigm, *Journal of Mixed Methods Research*, 4(1), 6–16.

Zeitlin, M. (1991) Nutritional resilience in a hostile environment: positive deviance in child nutrition, *Nutrition Reviews*, 49(9), 259–268.

Chapter Two: Positive Deviance, Big Data and Development: A Systematic Literature Review

Basma Albanna and Richard Heeks

Abstract

Positive deviance is a growing approach in international development that identifies those within a population who are outperforming their peers in some way; e.g. children in low-income families who are well-nourished when those around them are not. Analysing and then disseminating the behaviours and other factors underpinning positive deviance is demonstrably effective in delivering development results. However, positive deviance faces a number of challenges that are restricting its diffusion. In this paper, using systematic literature review, we analyse the current state of positive deviance and the potential for big data to address the challenges facing positive deviance. From this, we evaluate the promise of "big data-based positive deviance": this would analyse typical sources of big data in developing countries – mobile phone records, social media, satellite imaging, sensor data, etc. – to identify both positive deviants and the factors underpinning their superior performance. While big data cannot solve all the challenges facing positive deviance as a development tool, it could reduce time, cost and effort; identify positive deviants in new or better ways; and enable positive deviance to break out of its current preoccupation with public health into domains such as agriculture, education, urban planning and more. In turn positive deviance could provide a new and systematic basis for extracting real-world development impacts from big data.

2.1 Introduction

Many development practitioners continue to use a traditional "needs-based" approach to development, involving top-down identification of needs and problems, and the external imposition of solutions that meet those needs. This type of approach can work well in addressing specific technical challenges. But it works much less well where development requires learning and behavioural change by beneficiary groups; something which necessitates much greater knowledge of and engagement with beneficiary communities (Saïd Business School 2010; Pascale, Sternin & Sternin 2010; Singhal 2011; Nel 2017). As a result, more bottom-up "asset-based" approaches have come into existence, which capitalize on a community's inherent assets and capabilities – including knowledge – in solving development problems. Positive deviance (PD) is one such asset-based approach. It is based on the observation that in every group or community, a few individuals use uncommon practices and behaviours to achieve better solutions to problems than their peers who face the same challenges and barriers (Pascale, Sternin & Sternin 2010). Those individuals are referred to as "positive deviants" (PDs) and adopting their solutions on a wider basis is referred to as the PD approach.

The term "positive deviance" was first used in 1976 to describe a practical strategy for the design of food supplementation programmes in Central America; a strategy that was derived endogenously rather than exogenously through identifying dietary practices developed by mothers in low-income families who had well-nourished children (S. M. Wishik & Van Der Vynckt 1976). The results of this study were not widely publicized, limiting uptake. It was not until the 1990s that PD started to be seen as a credible strategy for nutrition research and action, based on an accumulation of evidence of impact (Zeitlin 1991; Sternin et al. 1997; Sternin et al. 1998). The 1990s also witnessed its first large-scale adoption in international development by Save the Children, which used PD as a strategy to reduce malnutrition in Vietnam, rehabilitating an estimated 50,000 malnourished children in 250 communities (Sternin 2002). But it has only really started to attract attention in the 2000s, when Sternin and collaborators promoted PD more broadly as an asset-based approach for social change and demonstrated how it can be operationalized across a variety of development domains (Sternin & Choo 2000; Sternin 2002). Since the early 2000s, PD has been applied across multiple development domains, with public health being the most prominent.

As will be discussed in further detail below, PD tends to rely on in-depth primary data collection in identifying PDs and then community mobilization in disseminating and scaling successful practices. Identification is therefore time and labour intensive, with costs proportional to sample size (Lapping et al. 2002a; Marsh et al. 2004; Felt 2011). As a result, PD has traditionally made use of relatively small-scale samples. Statistically and practically, this can make it harder to identify positive deviants, given their relative rarity (Marsh *et al.* 2004). It also limits the ability to accurately generalize the identified practices to larger populations (Marsh *et al.* 2004). Path dependency has also been evident in uptake of the PD approach, in terms of geographical distribution and domain of application, with most studies concentrated in a few countries of Asia and in addressing malnutrition: the region and domain where it was initially introduced and practised by Sternin et al.

Given these and other challenges and limitations, there are obvious opportunities for innovation in positive deviance. Our particular interest here is in the innovative opportunities offered by big data: the increasing amounts of data about what we are and what we do and what we say, generated from digital devices, which provide an opportunity to gather insights into human behaviour. If big data can provide insights into behaviour, then big data analytics could identify patterns of "abnormal behaviour": variances from the average collective behaviour of observed units which could include the behaviours of those which a PD approach would define as positive deviants.

In this paper, we therefore investigate the potential of "big data-based positive deviance" (BDPD). Our particular interest flows from the line of argument above: that there are challenges for the traditional positive deviance approach which big data might be able to address. But there is also a converse interest: that positive deviance might represent a new approach to the extraction of development value from big data.

To investigate this potential, we have undertaken a systematic literature review (SLR) of the empirical applications of positive deviance and of big data in developing country contexts, in order to answer three questions. First, how is PD currently being applied in development? In particular, we seek to identify from the literature challenges in that application which new approaches might seek to address. Second, how is BD currently being applied in development?

We investigate this particularly in light of the challenges to positive deviance identified earlier; but we also extract challenges in use of big data. Third, what development value might result from the combined use of BD and the PD approach? Here we combine the findings of both literature reviews to address the interests expressed above: how big data can address PD challenges, but also how PD might be a valuable approach to use of big data in development.

The paper begins with a brief description of the method used in conducting the literature review, followed by three sections that answer each of the questions in turn: a presentation of the findings from our PD and then BD literature review before concluding with a discussion about big data-based positive deviance.

2.2 Systematic Literature Review Methodology

2.2.1 Literature Search and Selection

A systematic review of the literature was conducted using an adaptation of the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) protocol (Moher *et al.* 2009). The review included academic, peer-reviewed, English-language literature that reported empirical results using secondary or primary data sources from developing countries. The literature search was implemented using Google Scholar because 1) it is free and easy to access, making the SLR reproducible; 2) both PD literature and BD literature are multidisciplinary so it was important to use a non-disciplinary comprehensive base of literature; and 3) Google Scholar has the widest coverage of academic articles in comparison to other search engines and databases (Khabsa & Giles 2014). The utilized search strings and strategy are summarized in Table 2[1].

[1] The first part of the search strings looks for terms in the title of a paper; the terms in brackets search within the text of the paper.

Positive Deviance Literature	
Search String	Str1: intitle:"Positive deviance" (study OR empirical OR practice OR experimental OR survey OR fieldwork) Str2: intitle:"Positive deviants" (study OR empirical OR practice OR experimental OR survey OR fieldwork) Str3: intitle:"Positive deviant" (study OR empirical OR practice OR experimental OR survey OR fieldwork)
Time Period	1970-2017
Exclude	Patents, Citations and non-English Results
Big Data Literature	
Search String	intitle:"Big data" ("developing countries"[2] AND application)
Time Period	1998-2017[3]
Exclude	Patents, Citations and non-English Results

Table 2: Google Scholar search strategy used for the literature search

To retrieve relevant studies, we used the **intitle** operator, which ensures that the title of the retrieved articles would include the words following the operator. We also used **AND** and **OR** Boolean operators to ensure the existence of key terms in the text of the articles, thus reducing the time required in screening irrelevant sources. For example, in the PD literature search, the words "positive deviance", "positive deviant" or "positive deviants" were used with the "intitle" operator; thereby targeting articles that have PD as the central theme. The words "study", "empirical", "practice", "experimental", "survey" or "fieldwork" were used for the in-text search to restrict articles not providing empirical evidence. The same search strategy was used to retrieve the BD literature but with a simpler search string found to deliver the corpus of

2 The term "developing countries" was not used in the PD search as initial investigation suggested that the majority of studies were in those countries, and we could then manually exclude those that were outside scope. Conversely, it was used in the BD search as the majority of BD literature was not in developing countries, and so the term was seen to be a useful means to quickly narrow the search to more-relevant items. As noted below, however, the term itself was in practice rather narrow and served to omit a number of relevant studies which then had to be manually identified and included via backward snowballing.

3 To the best of our knowledge, the term "big data" was first introduced in 1998 (Mashey 1998), thus setting the boundary for the search period.

literature suitable for analysis. To ensure that key studies were not excluded, backward snowballing[4] of relevant articles was employed and it led to the identification of one additional PD article and 22 additional BD articles[5]. A total of 75 articles were included in the final corpus of analysis: 41 PD articles and 34 BD articles. Figure 1 reports on the identification and selection protocol.

2.2.2 Content Analysis

NVivo was used for the qualitative and quantitative content analysis of the selected articles. For each article in the PD and BD corpus, the following attributes were identified and used for classification: title, year of publication, research methodology, research approach, types of data used, sample area (i.e. rural or urban), sampling unit, country, region and study duration (if stated). Those attributes were derived based on a mix of commonly-used data fields in systematic literature review and an iterative process of attribute selection depending on what arises as an important variable for the topic of analysis (Petticrew & Roberts 2005; Okoli 2015). Additionally, articles where coded into several nodes based on the areas covered in the qualitative analysis; those areas identified iteratively based on our overall purpose of understanding the potential development value of combining big data- and positive deviance-based analysis. Those areas can be summarized into: 1) challenges and limitations, 2) benefits, 3) conceptual frameworks, 4) methods and data, 5) research findings and 6) research opportunities. The following sections do not report all content analysis but only those main elements seen as relevant to the purpose of this paper.

[4] Backward snowballing involves screening the reference lists of relevant literature review articles to look for additional literature. It is considered an effective technique for conducting complex systematic literature reviews (Greenhalgh & Peacock 2005).

[5] The number of BD articles identified by backward snowballing was much greater because many relevant big-data-for-development articles either do not specifically use the words "big data" in their title (e.g. they use "mobile data" or "satellite images"), or do not specifically use the words "developing countries" in their text (e.g. they use the name of a particular country or a development goal).

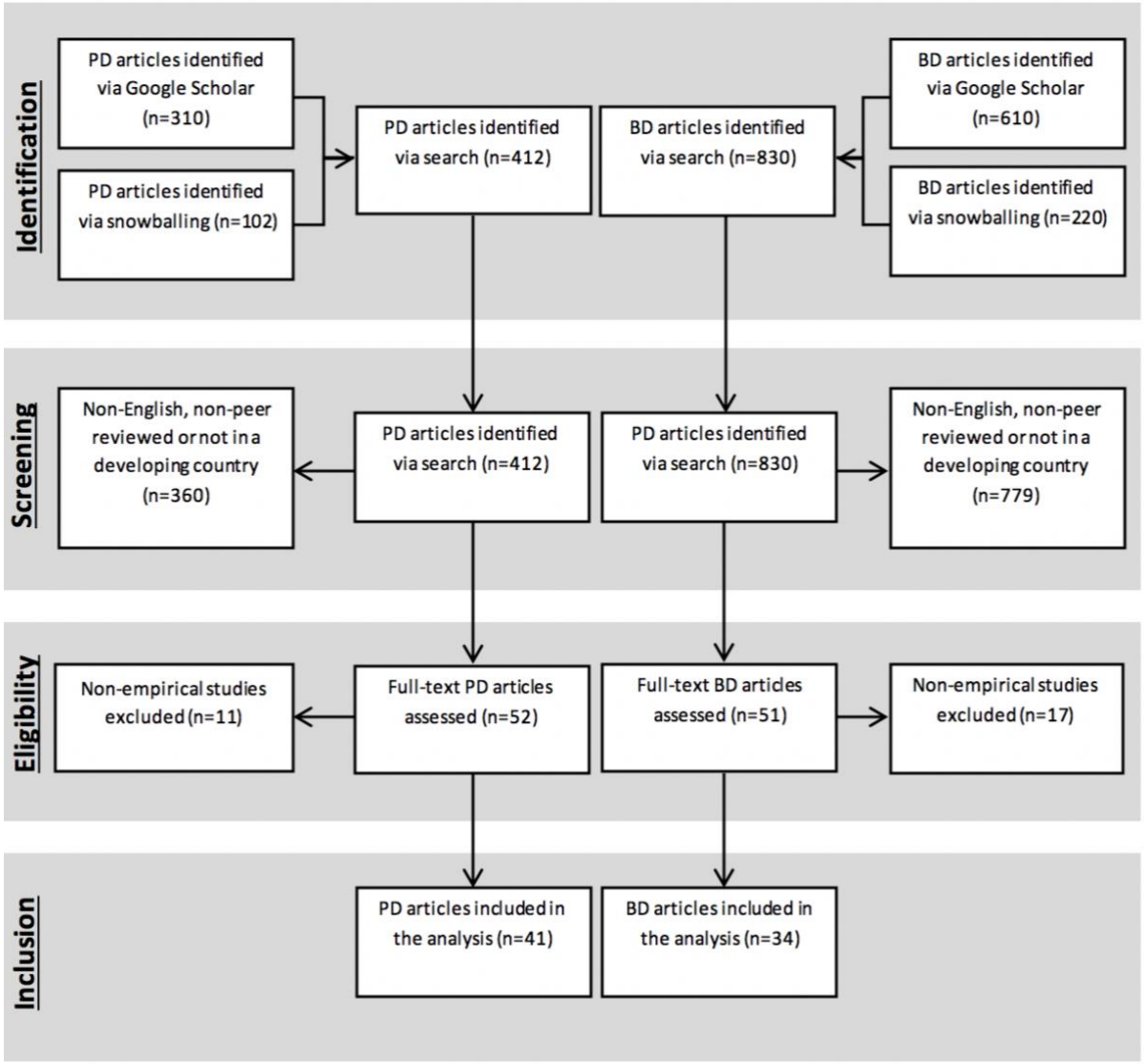

Figure 1: Flow diagram for identification and selection of PD and BD articles (adapted from the PRISMA protocol)

2.3 Positive Deviance

According to the Positive Deviance Initiative (Springer, Nielsen & Johansen 2016), the successful application of PD has been reported in more than 60 countries across the globe with a total outreach of more than 30 million individuals for the period between 1990 and 2016. Applications include: reducing childhood malnutrition, enhancing school retention, eliminating neonatal

mortality, limiting HIV transmission, improving salesforce productivity, fighting against female genital cutting, enhancing healthcare services, reducing transmission of antibiotic resistant bacteria in hospitals, and enhancing pregnancy outcomes. The central premise of the PD approach is that it harnesses the inherent wisdom of individuals existing within a community to develop solutions to their own problems. And since solutions come from the people, they take into account contextual and cultural variables, making them less vulnerable to social rejection. PD is also considered an efficient approach within international development, because it reduces reliance on aid and external expertise and instead capitalizes on local resources and know-how. It can also generate local engagement in identifying and disseminating practices and is seen as creating self-efficacy (individuals' belief in their capacity to execute behaviours necessary to achieve a desired objective): often considered to be a key influencer in the adoption of recommended behaviours (Babalola et al. 2002; Babalola 2007).

Much of the positive deviance literature has a – forgiving the pun – rather positive, even proselytizing tone. Balancing this, there are some more critical insights with three particular concerns being raised[6]. First, a concern that – compared to its practical application – the ideas of positive deviance lack conceptual clarity, with papers using different definitions and with limited theorization of positive deviance (Herington & van de Fliert 2018). Second, a concern that positive deviance does not always work in practice. Problems have included difficulties in identifying PDs (Marsh *et al.* 2004) and/or their differential characteristics and behaviours (Bradley et al. 2009; Felt 2011); inability to scale the PD solutions across a community and, particularly, between communities[7] (LeMahieu, Nordstrum & Gale 2017). These concerns overlap significantly with material on the third area of concern: practical challenges to the implementation of positive deviance; a topic discussed further below as an outcome of the SLR.

[6] Specifically, within the field of social psychology, there has been opposition to the idea of positive deviance by those working on deviant behaviour, who wish to solely assign a negative connotation to deviance. However, such arguments do not transfer beyond the specific field of deviant studies and have, in any case, been fairly well refuted (Shoenberger 2017).

[7] With an "orthodox" view of PD being that it should not seek to transfer solutions between communities, but develop them within each community (LeMahieu et al. 2017). Even if one accepts this view, it clearly depends on where one sets the definition and boundary of "community".

In sum, though, one may conclude that there has yet to be a weight of critique sufficient to discredit positive deviance as a development approach or to identify aspects necessary and inherent to PD that would undermine it. Conversely, there is a growing weight of evidence demonstrating beneficial development outcomes emerging from its application.

That application generally follows the five steps of the PD methodology, which can be outlined as follows (Positive Deviance Initiative 2010):

1. Defining the problem and determining desirable outcomes
2. Discovering PDs i.e. individuals or other social entities who unexpectedly achieved the desired outcomes
3. Determining the underlying practices that led to those outcomes (this is known as positive deviance inquiry (PDI))
4. Designing interventions to enable others to access and practice new behaviours
5. Monitoring and evaluating the PD intervention

Building from this background on positive deviance, the systematic literature review begins with a timeline showing the volume of PD literature over the last two decades followed by a thematic classification of the literature. We then analyse the secondary data sources used in previous PD studies as these may share characteristics with attempts to use big data in PD. We then discuss the different units of analysis used in the literature before presenting the identified challenges of the PD approach; those challenges presenting potential opportunities for big data to make a contribution.

2.3.1 PD Literature Timeline

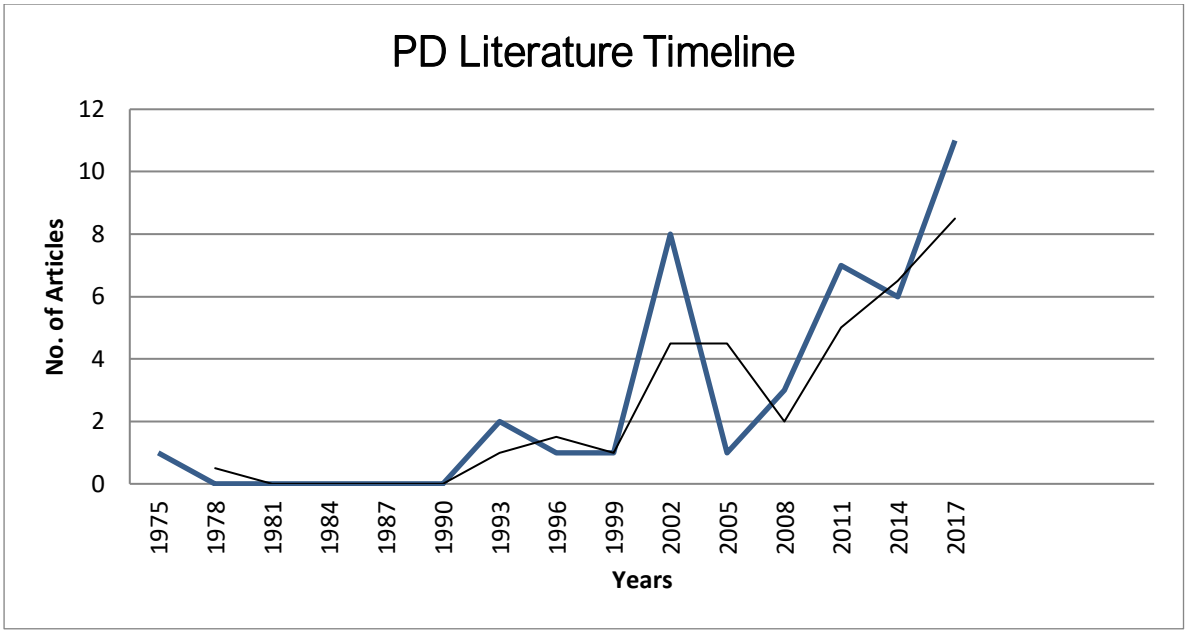

Figure 2: Timeline of reviewed PD literature for period 1976-2017

The publication timeline for work on positive deviance is shown in Figure 2[8], beginning with the first empirical PD study, published in 1976 (S. M. Wishik & Van Der Vynckt 1976). This article published the impending methodology but did not publish results, and it was not until the early 1990s that the approach started to gain attention due to the book (1990) and study (1991) published by Zeitlin which provided extensive observations on the PD approach in nutrition with a strong emphasis on impact. There was a peak in the early 2000s, which we can owe to Sternin who operationalized the PD approach and published the results of its application in the Save the Children project which reduced malnutrition in Vietnam by 65% to 80% in two years (Sternin & Choo 2000). This led to an extensive and rigorous evaluation of the PD strategy in solving child malnutrition, revealing positive results that supported its uptake in this field at this time (Hendrickson et al. 2002; Lapping et al. 2002b; Mackintosh et al. 2002). From the mid-2000s, there has been steady growth, with particular expansion in recent years: a possible explanation could be that in 2016, three international PD-focused conferences were held for the

[8] The blue line represents the actual number of studies whereas the black line represents a projection of the trend.

first time. Whatever the particular reason, it suggests growing interest and activity around positive deviance in developing countries; encouraging further work in this domain.

2.3.2 PD Research Approaches

Four research approaches to positive deviance were identified from the 41 reviewed articles, Normal PD (25 studies), Comparative PD (7 studies), Programmatic PD (6 studies), and PD Evaluation (3 studies). Below is a summary of each; detailed here so that the reader may understand better the type of development activity to which positive deviance has so far been applied and in what manner:

Normal PD: This is the most common approach, which applies the PD approach to a single group. Most studies of this type stop at the PD inquiry stage (i.e. step three of the PD methodology outlined above), where the uncommon, successful practices of PDs are identified, without going further into designing interventions to promote those practices and monitoring progress. For instance, in Lackovich-Van Gorp (2017), a study was conducted to investigate strategies that could prevent marriage by abduction in Ethiopia. PDs were girls over 18 years old coming from very poor households who were still not married. The intervention applied only the first three steps of the PD methodology to identify PDs and the strategies they employed to protect themselves from marriage by abduction. The average duration of such studies is eight months. 68% of those studies use mixed methods, 21% use quantitative methods and 11% used qualitative methods. The normal PD approach covered studies that tackled issues including: healthcare-associated infections (de Macedo et al. 2012; Marra et al. 2013), enhancing health outcomes of women in disadvantaged circumstances (Long *et al.* 2013), cancer prevention (Vossenaar et al. 2009; Vossenaar et al. 2010), child marriage (Lackovich-Van Gorp 2017), child rearing (Aruna, Vazir & Vidyasagar 2001), infectious disease control (Babalola et al. 2002; Babalola 2007; Nieto-Sanchez et al. 2015), improving pregnancy outcomes (Ahrari *et al.* 2002), counselling for family planning (Kim, Heerey & Kols 2008), child malnutrition (Wishik & Van Der Vynckt 1976; Shekar et al. 1991; Shekar et al. 1992; Guldan et al. 1993; Merchant & Udipi 1997; Bolles et al. 2002; Sethi et al. 2003; Kanani & Popat 2012; Aday et al. 2016; Roche et al. 2017; Merita et al. 2017), neonatal mortality (Marsh *et al.* 2002) and managing medico-social problems through self-care (Gidado *et al.* 2010).

Comparative PD: Studies in this research approach compare the results of two methodologies each applied on a different group, having PD as one of the methodologies. It includes control trial study designs where a PD intervention is applied to one group and the outcomes are compared with an equivalent control group that was not exposed to the PD intervention. As an example, in one of the reviewed studies (Lapping et al. 2002b), a PD inquiry was compared with a case-control study to identify factors associated with nutritional status of Afghan refugees in Pakistan (concluding PD to be at least as good if not more effective than control study in factor identification). Research in such studies is mainly mixed methods and sometimes it is only quantitative. Study durations are seven months on average. The comparative PD approach includes studies that tackled: healthcare-associated infections (Marra et al. 2010; Escobar et al. 2017), malnutrition (Hendrickson et al. 2002; Ndiaye et al. 2009; Nishat & Batool 2011) and clinical performance in medical schools (Zaidi *et al.* 2012).

Programmatic PD: Studies belonging to this approach aim at understanding why a few individuals (PDs) respond to a development intervention programme better than their peers who are targeted by the same intervention. The PD inquiry is used to identify reasons behind the successful responses of the PDs and the findings are used to inform intervention strategies and to increase overall adoption. For instance, in Garrett & Barrington (2013), a qualitative study was conducted to investigate barriers that prevent Honduran women from engaging in a cervical cancer screening programme. PDs were women that engaged in the uncommon but beneficial practice of screening. The PD intervention was designed to identify those women and the factors that led to their uncommon behaviour. Those factors (e.g. self-love and self-support) were to be used in future screening promotion efforts. Research in such studies is either quantitative or mixed methods. Study durations are on average three months long. And since programmatic studies' main interest is just in post hoc identification of reasons for deviants to adopt or engage with the focal programme, they usually end at the third stage of the PD methodology. Example studies include programmes concerned with malnutrition (Levinson et al. 2007; D'Alimonte et al. 2016; Sethi et al. 2017), farmer training (Tekle 2015) and livestock feed technology adoption (Birhanu, Girma & Puskur 2017).

PD evaluation: This is the least-prevalent research approach, which aims at evaluating the sustainability and impact of a PD intervention. Studies are also the longest with an average

duration of 18 months and rely mainly on mixed methods in evaluation. The reviewed literature included three evaluative studies in malnutrition (Lapping et al. 2002b; Mackintosh, Marsh & Schroeder 2002; Anino, Were & Khamasi 2015) and one study in infectious disease control (Marra *et al.* 2011).

2.3.3 Sources of Data

All the reviewed studies used primary data for PD identification and inquiry except for two studies that used secondary data. The first of this latter group was an exploratory study (Long *et al.* 2013) that investigated the factors associated with positive health outcomes among rural women in West Bengal. It used previous data from a randomized control trial conducted in a rural population, on 2,227 consenting women and adolescent girls. Using quantitative analysis only, it was possible to examine the characteristics of PDs and factors affecting better health outcomes. However, there was limited ability to examine other possible factors affecting the targeted outcome, since the tool used to collect data for the previous study was not designed for the same purpose as this later study. The second study (Birhanu, Girma & Puskur 2017) was also an exploratory study that aimed at investigating the factors leading to better adoption of livestock feed technologies in Ethiopia. It used a previous household survey that included 603 farm households and aimed at identifying successful cases of improved livestock feed technologies and factors underpinning this success. Since the original study had the same purpose as the PD study, the collected data was able to unveil all possible factors affecting the desired outcome through quantitative analysis. These studies indicate the potential to undertake positive deviant identification without a need for primary research; thus, signalling the potential for big data-based PD studies but also the challenge of re-purposing datasets not specifically gathered for PD purposes.

2.3.4 PD Unit of Analysis

The majority of the reviewed PD studies had individuals (infants, children, mothers, patients, students, healthcare workers, etc.) as their primary unit of analysis, except for three studies that investigated positively-deviant farmer training centres (Tekle 2015), farm households (Birhanu, Girma & Puskur 2017) and (disease-resistant) houses (Nieto-Sanchez *et al.* 2015). None of the

studies conducted aggregation analysis e.g. identifying community-level deviance instead of individual-level deviance. This can be attributed to the small sample size in terms of number of communities covered that would not permit the identification of this type of deviance; a limitation that larger-scale datasets might not suffer.

2.3.5 PD Challenges

Analysis of the literature on positive deviance reveals a series of challenges or limitations arising from work to date; challenges which we will later interrogate to see if big data might have some response:

Time and cost: The application of the PD approach is time-consuming (Lapping et al. 2002a; Marsh et al. 2004; Felt 2011). As can be seen from the data in Section 2.3.2, it takes months to complete the phases sequentially. Alongside concerns about the time requirements, are also concerns that the quality of implementation may be compromised due to time constraints. For instance, one of the studies reported that the desired large sample size was not obtained because of time limitations (Nishat & Batool 2011). Since PD depends typically on primary data collection, community participation, face-to-face interviews and observation, the cost of PD interventions also tends to be high. As with the time constraint, cost is also a function of sample size, which can encourage smaller samples. In addition, collecting primary data from some high-risk areas brings with it additional time, cost and complexity in order to mitigate the risks (Shekar, Habicht & Latham 1991).

Positive deviant identification: Within any given population, positive deviants are relatively rare. Based on those reviewed studies that provide the necessary data, we can calculate an average prevalence rate of 11%. This is slightly higher than, but not completely out of line with, earlier estimates that PDs typically form 0-10% of a population (Marsh *et al.* 2004). Whatever the exact figure, PDs are statistical outliers, and sample size thus plays a role. As the sample size increases, the more representative of the population it becomes, and thus the likelihood/prevalence of positive deviants becomes greater (Osborne & Overbay 2004).

Hence, there is a statistical pressure to undertake large sample size studies in order to identify a sufficient sample size of PDs. However, given the time- and labour-intensity of PD just noted,

with costs proportional to sample size, there is a counter-pressure to keep overall sample sizes small. For example, in the comparative study of Afghan refugees in Pakistan (Lapping et al. 2002b) the compared groups were 8 and 50 strong. Another study in Egypt that addressed factors associated with successful pregnancy outcomes, reported that the information gained from PDs was limited; this can be attributed to their very small sample size (n=11) (Ahrari *et al.* 2002). Similarly, in the Honduras study examining women who overcame barriers to cervical screening, the sample size (n=8) was seen as not large enough to achieve full saturation of relevant factors. The use of very small samples for PDI not only potentially misses important aspects of PD behaviour, but would also have less statistical power to identify valid associations. Additionally, as previously mentioned in Section 2.3.4, small sample size limits the ability to identify deviance at different levels of aggregation. One potential solution would be the use of large secondary datasets, which could be analysed at low cost while not compromising the number of PDs identified.

Moreover, PD primary data collection – often due to its time and cost – is undertaken via a cross-sectional not longitudinal design. It provides a single snapshot of the population since it depicts the behaviours of the analysed units at a certain point of time; hence, deviance becomes static and could be referred to as *point anomaly* in statistics (Goldstein & Uchida 2016). What it cannot do is identify the dynamics of deviance such as *contextual/conditional anomalies* that arise due to the particular condition of a context, with those conditions potentially differing over time. For example, one of the reviewed studies sought to identify preventive measures to control Chagas disease; PDs were bug-free houses throughout the period of inspection. However, an identified limitation of the study was that the houses selected are not necessarily bug-free throughout the year, since the entomological searches happened during the summer, and natural factors could have affected the results (Nieto-Sanchez *et al.* 2015). Hence, a few of those PDs might have been false positives: appearing as a point anomaly attributed to the individual house but in fact a contextual/conditional anomaly. Again, one potential cost-efficient solution would be large secondary datasets; in this case, where the data was collected longitudinally.

Methodological risk: Alongside the practical risks of PD given the need for large-scale primary fieldwork in developing countries, we were able to identify two methodological risks associated

with use of the PD approach. First, there is a PD behaviour identification risk. For example, some of the studies (Marra et al. 2010; de Macedo et al. 2012; D'Alimonte et al. 2016) that used observational methods in PD inquiry, reported the potential for a *Hawthorne effect*: an alteration of the behaviour of the subjects of a study due to being observed. Another risk is the inability to extract successful strategies and behaviours practised by the positively deviant individuals. Positive deviance methods presume the willingness of PDs to share their strategies and best practices. However, this might compromise what the deviants see as a competitive advantage over others resulting from their outlier behaviour, leading them to be unwilling to share (Felt & Cody 2011). For example, in one of the reviewed studies (Zaidi *et al.* 2012), positive deviance was used to try to identify and disseminate the strategies employed by successful medical students, in order to improve the clinical performance of their peers. There is a potential risk that the high performers would refrain from sharing their best practices when interviewed. In both cases, analysis of behaviour via secondary / remote observation could help to avoid these risks.[9]

Second, there is a risk of not being able to establish a cause and effect relationship between PD interventions and achieved results (step 4 of the PD methodology). Some studies (e.g. Nishat & Batool 2011, Nieto-Sanchez et al. 2015) noted that results could not be attributed to the PD intervention alone, since the targeted population might have been exposed to other interventions and external factors that might have contributed, partially, to the desired outcome. This challenge of attribution (and also issues of time and cost) may be one explanation behind the limited number of evaluative studies of PD interventions: i.e. those that moved to step 5 of the PD methodology (Ndiaye et al. 2009; Roche et al. 2017). (Another explanation may be the lack of guidance on how to apply credible monitoring and evaluation techniques (Lapping et al. 2002b; Felt 2011).) As noted in Section 2.3.2, only 7% of the reviewed studies were evaluative: very low considering the importance of understanding the development impact and

[9] Though noting as per the point raised in Section 2.3.3, that relying on quantitative analysis of secondary data to infer practices might limit the ability to test the effect of other potential factors influencing the desired PD outcome, especially in studies where the instruments used to collect the data were not designed to measure the desired outcome (Long et al. 2013).

value of PD approaches. Being able to demonstrate the lasting success of a PD intervention could support its wider adoption, and the wider adoption of PD more generally.

Scalability: There are two challenges underlying the scale-up of PD interventions. The first challenge is in scaling practices within a community. PD relies heavily on community engagement to promote the adoption and mobilization of the identified practices and to achieve behavioural change through self-efficacy. For instance, 25% of the reviewed studies employed the *PD Hearth* (Wollinka *et al.* 1997); a nutrition education framework designed to empower mothers to enhance the conditions of their malnourished children. It requires mothers of PD children to host neighbouring mothers and their malnourished children for 12 consecutive days, where they prepare meals and feed their children together (Lapping et al. 2002a; Marsh et al. 2004; Pascale, Sternin & Sternin 2010; Felt 2011). With this level of engagement, PD proved successful in small-scale adoption, but made large-scale adoption a challenging task given both the growing complexity but also the challenge of a strong enough pre-existing social fabric to ensure cooperation of the PD mothers. (Zeitlin (1991) notes a similar point that – for certain communities and domains of action – there could be resistance to making everyone a "top performer"; and that moves towards that might disrupt and disintegrate system dynamics.)

The second challenge is the scaling of practices across communities. An issue with PD is the inability to generalize practices and behaviours inferred from one community to another. In the majority of the reviewed studies, PD interventions targeted small-scale communities, and the inferred practices were particular to the circumstances of this community making it difficult to replicate in other communities (Saïd Business School 2010). If broader, cross-community data could be accessed – identifying PDs and their behaviour on a wide scale – then this challenge of limitations on generalization could be reduced to some extent; though of course this would likely assume/require the presence of non-community-specific behaviours underlying positive deviance.

Narrow domain/geographic scope: There is a current skew in the domain and geographic focus of PD applications in developing countries; summarized in Figure 3. Regarding domain coverage, we found that the vast majority – 89% – of the reviewed studies were in public health, with 41% focused specifically on malnutrition. Put another way, there were only four non-health

studies: two on agriculture (Tekle 2015; Birhanu et al. 2017), one on child protection (Lackovich-Van Gorp 2017) and one on education (Zaidi *et al.* 2012). As for the geographic coverage, there are nearly 150 developing countries (OECD 2017) but PD studies identified by the review encompassed only 20, with just four countries (India, Brazil, Pakistan and Ethiopia) responsible for almost 50% of studies, and only two countries having hosted studies from more than two domains. There is also a within-country geographic concentration, with 83% of studies being undertaken with rural communities; significantly out of kilter with population distribution in developing countries.

This domain and geographic concentration can be attributed to a form of path dependency in positive deviance. The first use of PD (S. M. Wishik & Van Der Vynckt 1976) was for a nutrition-based intervention in a rural community. Then, early adopters of PD who set the foundation for the field (Zeitlin 1991; Zeitlin et al. 1994; Sternin, Sternin & Marsh 1997) including development of an operationalizable framework for PD (Monique Sternin, Sternin & Marsh 1998) all undertook work on malnutrition in rural areas. Despite applicability of PD across many domains, subsequent PD actions have often followed suit, leaving a gap of domains and locations that have been largely ignored by PD to date.

That PD has relevance to other countries and other domains can readily be seen from its application in the global North e.g. to public sector reforms (Andrews 2015), enhancement of prison conditions (Awofeso, Irwin & Forrest 2008), organizational scholarship (Mertens *et al.* 2016) and waste management (Delias 2017). But PD for developing countries needs encouragement to spread further than its current narrow path

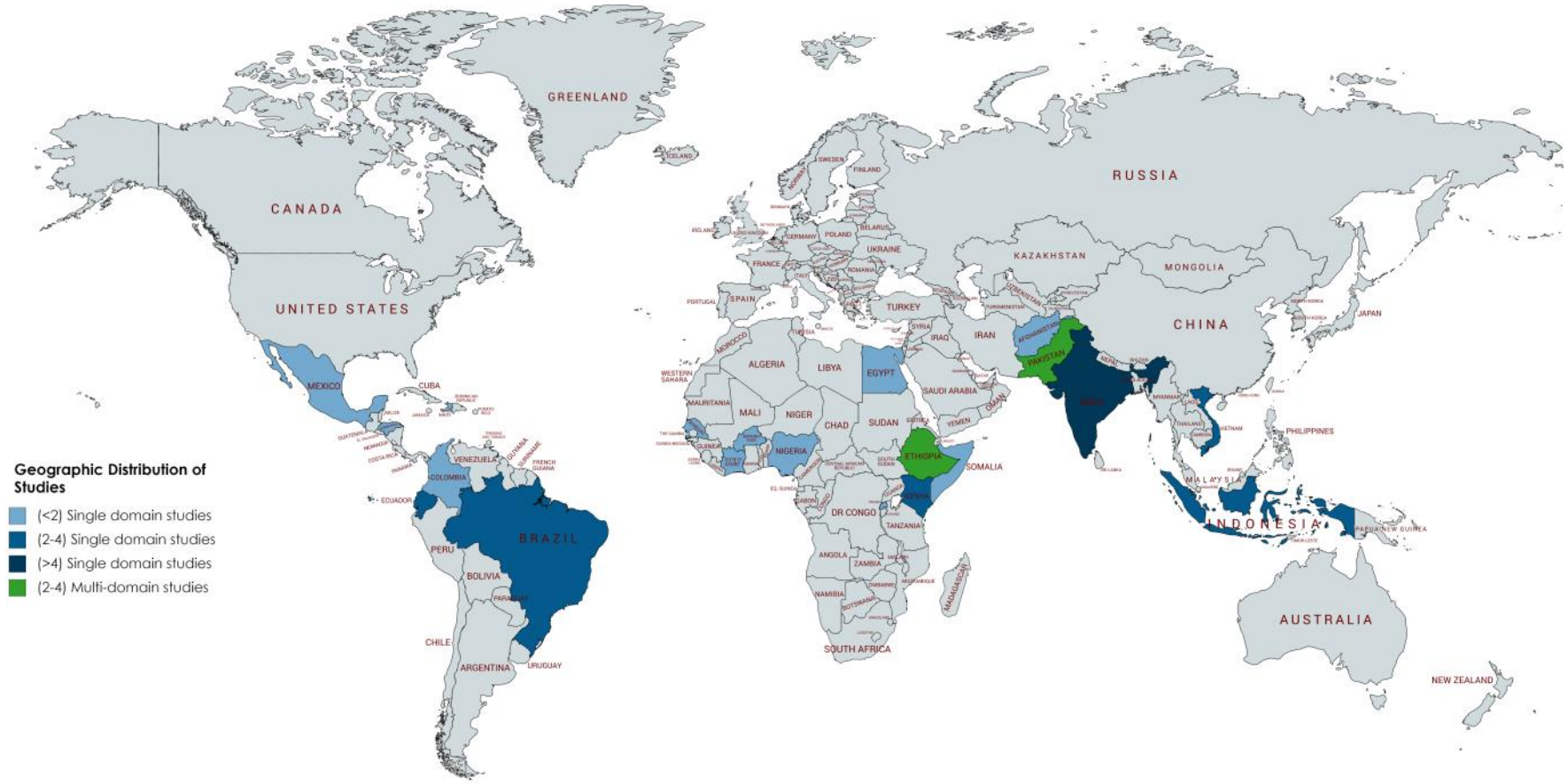

Figure 3: Geographic distribution of the domains of PD literature relating to developing countries

2.4 Big Data for Development

The evolution and diffusion of digital infrastructures has led to a proliferation of data, often referred to as "big data" (BD). There has also been an increasing ability to make use of it, characterized in enhanced processing and storage capacities, which has provided an opportunity to convert this data into information and knowledge that feeds decisions and actions. The main characteristics of BD derived from Gartner's definition (Gartner 2013) are: 1) Volume: huge amounts of data generated from the rapid diffusion of mobile phones, social media, and other online services, plus the growing use of sensors and satellite imagery; 2) Velocity: the growing speed and currency of data production, enabling decisions and actions to be taken in a timely manner; and 3) Variety: the combined availability and potential use of structured data, having a predefined structure (e.g. mobile transactions), and unstructured data, not having a predefined structure (e.g. email, video and audio), to extract insights.

The applications of "big data for development" (BD4D) span a wide variety of domains and leverage new sources of data and new analytical tools. It is argued that big data can fundamentally shift the way we pursue social change as it is capable of providing snapshots of the wellbeing of populations at high frequency, high degree of granularity, and from a wide range of angles, narrowing both time and knowledge gaps (UNGP 2012). Sitting alongside concerns about and critiques of BD4D (e.g. Taylor & Broeders 2015), big data is therefore also argued to offer new opportunities for development for reasons that include:

1) **Low cost:** Digital traces produced from digital platforms provide a low-cost alternative to traditional sources of data (e.g. censuses, surveys); in some instances, by replacing variables of interests with correlated proxies (Hilbert 2016). For instance, mobile call duration and frequency have been correlated to income or education levels in a geographic region (Frias-Martinez & Virseda 2013), and could then substitute for survey and similar data gathering.
2) **Real-time feedback and awareness:** Through monitoring populations, BD makes it possible to understand where policy and programme interventions are succeeding or failing in real time, in order to make adjustments in a timely manner.

3) **Broad sampling:** With a global average penetration of 95% and a 75% penetration among base of the pyramid populations (Cartesian 2014), mobile phones are coming close to sampling the universe N instead of sampling n of the universe N (Hilbert 2016).
4) **Detail and insight:** The ability to merge and use different sources of data reflecting a certain event or reflecting the behaviours of an individual, community or an organization provides a real-time, cross-validated, fine-grained picture of reality.
5) **Big data analytics:** Advanced analytics techniques – those which perform particularly well when applied to huge datasets – enable big data to be used for better decision-making. An example is machine learning, a subfield of artificial intelligence, which gives computers the ability to learn from data without being explicitly pre-programmed on the knowledge they will extract.

Given this potential value of big data to development, the review of literature outlined in Section 2 was undertaken, and is reported next. This section begins with a timeline showing the volume of BD literature over the last decade followed by a thematic classification of the literature and the geographic distribution of its application domains. We then discuss the other forms of data that were combined with BD, the different units of analysis and the employed BD analytics techniques before outlining the challenges of BD4D at the end. All this provides the knowledge of big data necessary to understand how it might be applied to action research on positive deviance; as then discussed in Section 2.5.

2.4.1 BD4D Literature Timeline

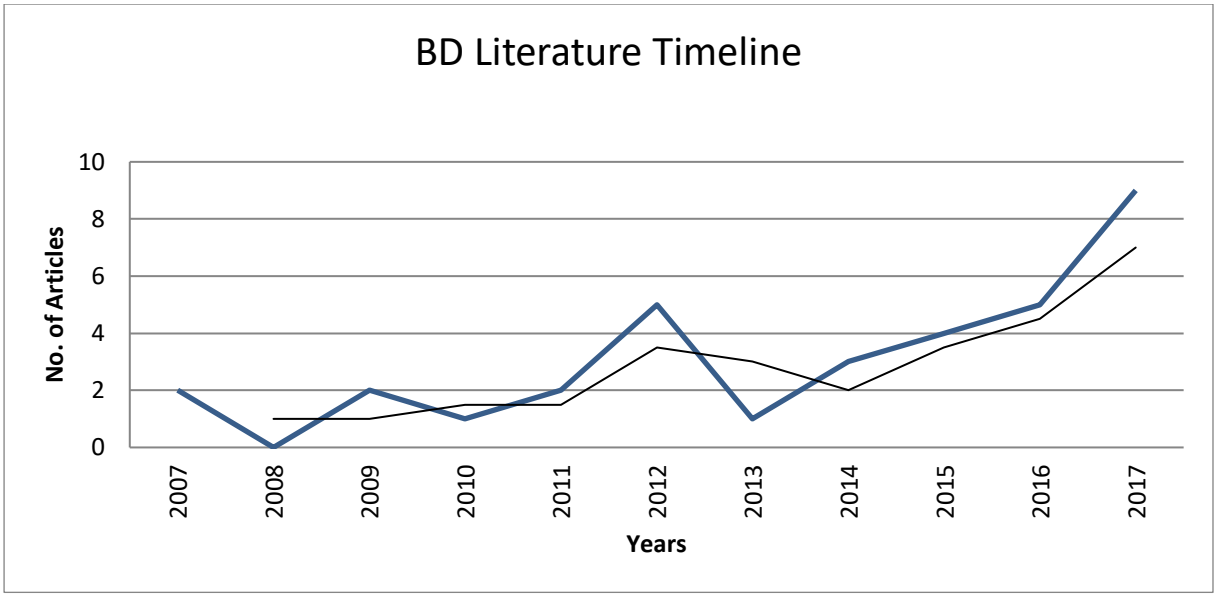

Figure 4: Timeline of reviewed BD4D literature[10]

2.4.2 BD4D Research Approaches

Utilising the taxonomy of BD applications proposed by Hilbert (2016), we classified the reviewed literature into four main approaches based on the elements being tracked: locations (12 studies), words (4 studies), nature (6 studies) and economic activity (12 studies). This gives a sense of the general scope of big data application; for example, its potential application in positive deviance analysis.

Tracking locations: This approach contains applications that analyse human and object mobility data. This typically comes from mobile phones in the form of de-identified call detail records (CDRs), which usually provide the time and associated cell tower of text messages and calls. This is the most common data type in the reviewed BD4D studies, notwithstanding the concerns of mobile phone operators about releasing such data: either that it has commercial sensitivity and could give competitors an unfair advantage, or that techniques could be used to de-anonymize the data and uncover individual subscribers' identities. There are also inherent

[10] The blue line represents the actual number of studies whereas the black line represents a projection of the trend.

biases in CDRs due to socio-economic and geographic variations in phone ownership, but evidence still suggests that CDRs provide the best description to date of population movement in low and middle-income countries (Wilson *et al.* 2016). CDRs have thus been used to analyse travel and migration patterns of mobile users to understand the spread of infectious diseases in low-income settings (Tatem et al. 2009; Bengtsson et al. 2011; Buckee et al. 2013; Wesolowski et al. 2014; Wesolowski et al. 2015); and to identify population displacement following disasters (Bengtsson et al. 2011; Lu, Bengtsson & Holme 2012; Wilson et al. 2016) and migration patterns in climate-stressed regions (Lu *et al.* 2016).

But CDRs are not the only location-related data that has been used. For example, data from a web mapping service application (Baidu.com) was used to calculate average travel distance to healthy food outlets in order to identify urban areas with limited food accessibility in China (Su *et al.* 2017). Car GPS data has been used to map the spatio-temporal distribution of pollution emissions from traffic (Luo et al. 2017; Huang et al. 2017). There are also studies that combine mobility data with other sources of data for better representation, cross-validity, data enrichment and for covariance analysis. For instance, in Tatem et al. (2014) physical data in the form of satellite images, climate and topographic data, was combined with CDR data to understand the spread of malaria; specifically, the seasonality of movements including movement across borders.

Tracking words: This approach contains applications that analyse actions, activities and events based on words, which typically come from social media. It usually faces the challenge of representational validity in terms of the demographic, socio-economic and geographic profile of contributors given skews in terms of those who do and do not use social media. There is also the challenge of potential differences between digital and real behaviour such as self-censorship or presenting a false image. One advantage, though, is that in many cases, data sources are readily accessible because they are "open data"[11] in nature; and they also benefit from enabling mapping of behaviour in real time (Pfeffer, Verrest & Poorthuis 2015). Examples of applications

[11] Data "freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control" (WP 2018).

include the use of Google search word trends to compare the demand for massive open online courses (MOOCs) between different countries (Tong & Li 2017), applying co-word analysis to map the research themes of Indonesian scholars' publications (Surjandari *et al.* 2015), analysing protest activity using twitter data during the Egyptian 25 January 2011 revolution (Wilson 2011) and revealing geographical and social patterns of tweets pertaining to flooding and criminal activity in Caribbean cities (Pfeffer, Verrest & Poorthuis 2015).

Tracking nature: This approach contains applications that use data to observe environmental and natural phenomena to mitigate risk, improve emergency response or to optimize performance (Hilbert 2016). Satellite imagery is the most common BD type used by those applications; this is due to its increasing availability at global and lower scales, often via open or other no-cost access. Growth in datasets over time is also allowing use for time series analysis. The reviewed studies falling under this approach used satellite imagery to detect illegal deforestation activities (Burgess *et al.* 2012), to monitor coal fires (Jiang *et al.* 2017), to map temporal water surfaces (Haas, Bartholomé & Combal 2009) and to model crop growth (Tesfaye *et al.* 2016).

Other studies collect environmental data via sensors. For example, sensor networks were used to monitor the spatio-temporal distribution of greenhouse gas emissions in China (Tang, Yang & Zhang 2014). There is also a study (Zhang *et al.* 2016) that combined sensor data, satellite images and meteorological data with social media for the analysis of urban waterlogging disasters (where drainage systems are unable to cope). Physical data was used to observe and understand waterlogging, and social media data (i.e. tracking words) was used to identify qualitative features of waterlogging incidents.

Tracking economic activity: This approach contains applications that use data that reflect the economic situation of the analysed units. Satellite images and CDRs are the two most popular BD sources that are used for this purpose. For instance, CDRs were used to predict wealth of individuals by tracking their history of mobile use represented in the intensity, volume, time or direction of calls (Blumenstock, Cadamuro & On 2015) or through phone ownership (Blumenstock & Eagle 2012). Similarly, satellite images were used to predict poverty and estimate economic growth in a number of studies. For instance, daytime satellite images were

used to predict socio-economic wellbeing by analysing visible household assets (Jean *et al.* 2016). They were also used to map population settlements (Tatem *et al.* 2007) and classify slums (Kohli *et al.* 2012).

Other studies used satellite nightlight images as a proxy for electricity consumption levels which, in turn, can be seen as a proxy for levels of economic activity (Sutton et al. 2007; Doll & Pachauri 2010; Henderson et al. 2012). However, nightlight images have difficulty distinguishing between poor densely-populated areas and wealthy sparsely-populated areas (Jean *et al.* 2016). Hence, Njuguna & McSharry (2017) complemented satellite nightlight data with CDR data to build a stronger poverty proxy by incorporating level of mobile usage. Other examples include use of data from a leading online retail platform in China (sofang.com) to analyse spatio-temporal features of housing prices (Li *et al.* 2017), and use of data from China's industrial enterprise database and customs import/export trade database to examine the extent of "greening" of global value chain enterprises (Song & Wang 2017).

2.4.3 BD4D Studies Domain/Geographic Distribution

Table 3 summarizes the overall domains and specific application topics of the reviewed BD4D studies. We can see that economics, public health and environmental studies were the most common application domains, within which the most common applications were infectious disease (16%) and poverty measurement (16%). In infectious disease studies (Tatem et al. 2009; Bengtsson et al. 2011; Wesolowski et al. 2014; Wesolowski et al. 2015), mobile data and health data are combined to identify risk areas. As for poverty measurement (Blumenstock et al. 2015; Jean et al. 2016; Njuguna & McSharry 2017), satellite nightlight and mobile data are used as a proxy for economic activity at temporal and geographic scales for which traditional data are of poor quality or unavailable.

Domain	Application	No. of studies
Economics	Predicting poverty	6
	Measuring economic growth	3
	Mapping population distribution	1
	Analysing housing prices over time	1
Public Health	Infectious disease control	6

Environment	Analysing traffic pollution	2
	Monitoring deforestation activities	1
	Monitoring changes in small water surfaces	1
	Monitoring greenhouse gas emissions	1
	Coal fire suppression efforts	1
Disaster Response	Population displacement post disasters	4
Food & Agriculture	Accessibility of healthy food stores	1
	Drought tolerant crops	1
Energy	Predicting electricity demand load	1
	Green technology adoption	1
Education	Quantifying MOOC demand	1
	Research theme mapping	1
Urban Governance	Urban waterlogging	2
	Security	1
Politics	Analysing twitter activity in revolutions	1

Table 3: Classification of BD4D studies by domain and application[12]

Figure 5 shows the geographic distribution of BD studies by domain; indicating that China had the biggest share of BD4D studies that span multiple domains. In addition, there were three studies not presented on this map since they were applied across multiple countries. The first study (Doll & Pachauri 2010) used night-time satellite imagery to estimate rural populations without access to electricity in developing countries. The second study (Henderson, Storeygard & Weil 2012) used satellite data to augment official income growth measures of coastal areas in sub-Saharan Africa. The third study (Haas, Bartholomé & Combal 2009) used remote sensing data to map temporary water bodies in regions of western sub-Saharan Africa. For within-country studies, some are rural-focused, many cover both urban and rural areas, and some are solely urban-focused.

[12] The total number of studies listed in the table is 37 although the reviewed BD studies were only 34. This is because two studies presented applications in multiple domains.

While the geographic coverage of big data studies in terms of countries is, as yet, not much better than that of PD studies, there is clearly ready potential for a much broader scope given the universal presence of at least some types of big data; including scope for more urban-focused work. And, in relation to domains, big data has already shown application to a wider range of topics than positive deviance.

Figure 5. Geographic distribution of the domains of BD4D literature

2.4.4 Sources of Data

37% of the reviewed studies complemented big data with other sources of primary and secondary data. For instance, in Sutton et al. (2007), a model was developed using GDP and population data, and night-time satellite imagery to predict GDP at sub-national levels. Similarly, Jean et al. (2016) used daytime satellite images, annotated with geo-referenced household consumption survey data, to develop a transfer learning model which trained the data in survey-rich countries to predict consumption and assets in survey-poor countries, using daytime satellite images alone. Wesolowski et al. (2014) demonstrated that community surveys can complement mobile data to approximate travel patterns of non-subscribers in rural areas. Secondary data is also used for cross validation, for example, in Wesolowski et al. (2015), where population-based surveys were used to validate the results of mobile data analysis that measured the magnitude of population displacements.

A number of studies combined health data, having geo-referenced disease cases, with mobility data to develop risk maps for infectious diseases (Tatem et al. 2009; Bengtsson et al. 2011; Wesolowski et al. 2014; Wesolowski et al. 2015). Spatially-referenced primary survey data has been used to support satellite imagery and mobile phone data in understanding the reality and the impacts of socio-economic or socio-spatial differences. For example, Blumenstock et al. (2015) was able to derive insights on the degree of wealth of individuals by supplementing mobile phone history big data with data from an anonymized mobile phone survey. Hence, it was possible, through training models, to predict wealth using only mobile phone use history for individuals not included in the survey sample. We may conclude from the above examples that – notwithstanding the potential for big data to provide a faster, cheaper and a more granular alternative to traditional data sources – greater value may be captured when BD is combined with those data sources instead of simply replacing them. In particular, big data can be validated in locations where comparator data exists, and then applied alone in locations where comparators are absent; an especially-helpful approach in developing countries where comparator data may be thin on the ground.

2.4.5 BD4D Unit of Analysis

In the reviewed BD4D studies, the unit of analysis ranged from individuals (e.g. mobile users, twitter users), through geographic areas (e.g. grid areas located via satellite imaging), to regions and even countries. The majority of studies applied aggregation on different scales to analyse and visualize patterns, events and spatial relations. Conversely, BD4D studies also provided a disaggregation opportunity. For example, in poor countries, data about economic growth is often available only at high levels of geographic aggregation (e.g. national level) because it is collected using sample surveys instead of location-disaggregated census surveys (Chandy, Hassan & Mukherji 2017). Due to the finer granularity of big data represented in user CDRs or night-time satellite images, it was possible in a number of studies (Sutton et al. 2007; Doll & Pachauri 2010; Henderson et al. 2012; Blumenstock & Eagle 2012; Blumenstock et al. 2015) to use those forms of data as proxies for economic growth at sub-national level. Compared to typical PD data, big data therefore may provide much greater aggregation and disaggregation potential.

2.4.6 BD4D Analytics

Discussion of big data in development can tend to focus on inherent qualities of the data such as the 3Vs: volume, velocity, variety. But greater value – including value for positive deviance – may vest in the advanced analytics techniques that are being applied to big data, and to use of these techniques for improved development decision making. The reviewed BD4D literature utilized two types of analytics:

1) **Descriptive analytics** provides information about the past and present. It uses data aggregation and data mining techniques to summarize historical data and answer the questions, "What has happened?" or "What is happening?". BD visualization is the essence of this type of analytics, creating a new face to standard descriptive statistical methods. For example, Pfeffer et al. (2015) used descriptive analytics to geo-map the word frequency of twitter data relating to two Caribbean cities which referred to crimes and flooding in order to better understand those phenomena.

2) **Predictive analytics** uses statistical models and forecasting techniques to answer the question, "What will happen?". It encompasses two types: inference and forecasting. Inference models predict the value of a certain variable of interest based on its association

with variables in another data source. For example, Jean et al (2016) used daytime satellite images and households' surveys, covering a specific area, to train a predictive model that was able to estimate the economic well-being, in another area, using only satellite imagery. This approach was able to overcome the data sparsity issue, through training models in data-rich areas to make predictions in data-poor areas. Inference models were also used to predict water precipitation and potential for waterlogging based on climate data and road and terrain maps (Zhang *et al.* 2016). On the other hand, forecasting models utilize trend analysis and pattern recognition techniques to predict what will happen in the future, based on what happened in the past. For example, in Ifaei et al. (2017) multivariate dynamic models were used to forecast power consumption using previous data on exported and imported power, and quantity of stored power over a period of time.

Big data also enables the use of intelligent data analytics techniques that perform better when applied to huge amounts of data. As noted above, an example is machine learning (ML). From the reviewed BD4D literature, ML techniques can be grouped into two main categories: supervised and unsupervised learning. In the former, ML is applied on a training dataset where each input X is labelled to a class or output Y and the primary objective of the learning algorithm is to develop a mapping function $Y=f(X)$, so that when you have any input x, you can predict its output y. For example, supervised learning was used to predict poverty levels from mobile phone data (Blumenstock, Cadamuro & On 2015) and from satellite imagery (Jean *et al.* 2016). In unsupervised learning, ML is applied on a dataset where you only have input data X and no corresponding output Y. The primary objective of the learning algorithm is to discover underlying similarities between the input data points and create clusters of data based on the perceived similarity. It is also capable of allocating new inputs into the appropriate cluster. For example, in one of the studies, unsupervised learning was used to find the interrelationship among academics' research approaches and cluster them, based on the co-occurrence of the publications' keywords (Surjandari *et al.* 2015).

2.4.7 BD4D Challenges

While there are broader challenges relating to the use of big data in development – such as associated shifts in power between different groups (Taylor & Broeders 2015; Sengupta et al. 2017) – there are also a set of more practical challenges that emerged from the literature

reviewed, which would need to be taken into account if using big data to research positive deviance.

Absence of theory: The majority of BD applications do not use theory-driven models, especially in cases of predictive analytics where they depend mainly on past data to predict what will happen in the future. However, attribution analysis (cause and effect studies) investigating why outcomes change in response to variations in inputs will need a theoretical framework. Employing a theory of change can guide the identification of explanatory variables (inputs) and indicators for outputs, outcomes and impact (Bamberger 2016).

Proof of concept skew: As might be expected given the relatively formative nature of big-data-for-development, most of the BD4D literature represents a proof of concept rather than use of data for actual development-related decision-making and implementation. As just one example, Pfeffer et al. (2015) demonstrate what (relatively little) tweets might tell us about location of urban flooding and crime but without any engagement with real-life urban planning decisions.

Representational validity: Mobile phone ownership is skewed towards certain population groups based on income, gender or age, leaving specific groups and geographic areas under-represented in mobile-based sources of big data (Bengtsson et al. 2011; Tatem et al. 2014; Wesolowski et al. 2014; Wilson et al. 2016). This is even more of a challenge for social media data, in terms of demographic and socio-economic profiling and the geographic spread of the content generators (Pfeffer, Verrest & Poorthuis 2015). BD sources were not particularly produced to investigate, assess or measure any of the presented development-related applications; they are rather a side effect (Pfeffer, Verrest & Poorthuis 2015). They provide one or more aspects of the studied issue, but they might overlook other important aspects requiring on the ground, targeted inquiry (Pfeffer et al. 2015; Lu et al. 2016). This explains the recent debates (Graham & Shelton 2013; Pfeffer et al. 2015) around the combined use of BD and other sources of data for better representational validity.

Human capacity: Both researchers and practitioners typically lack the necessary technical skills needed to clean up data sources, to link different data sources, to analyse big data, to identify emerging patterns from big data, and so on (Pfeffer, Verrest & Poorthuis 2015). There

are also highly unstructured data types, like satellite imagery, the analysis of which requires knowledge of advanced analytics and machine learning tools (Jean *et al.* 2016). Those missing skills and local conditions, especially in developing countries, limit the exploitation of this valuable data source, creating new digital divides (Batty *et al.* 2012).

Data accessibility: Most big data is not open and easily accessible. Data gatekeepers, such as mobile operators and public institutions, are not always willing to share their data; often because they consider it to be a source of commercial or political advantage (Pfeffer et al. 2015; Jean et al. 2016a).

Privacy and legal issues: It can be difficult to link and analyse different data sources while respecting privacy e.g. of individuals who produced the data (Sutton et al. 2007; Blumenstock et al. 2015). This can be particularly challenging in developing countries, where there is an absence of legal frameworks protecting citizens (Pfeffer, Verrest & Poorthuis 2015). As noted above, one reaction of data providers, like mobile operators, is to restrict access to datasets. Another reaction is to anonymize CDRs by removing and aggregating some attributes. While understandable, this can reduce the developmental value that can be captured from the provided datasets.

In summary, and despite the demonstrated value of using big data in a variety of development-related applications, it is important to note the challenges associated with its use. Of particular relevance for this paper are challenges that could affect the significance of its use in positive deviance, like privacy and accessibility. For instance, BD depicting human behaviour is the most relevant data for PD; however, if this data might compromise the security or privacy or undermine competitive advantage of data owners, its accessibility and usage would require strict rules and principles backed by adequate tools and systems to ensure "privacy-preserving analysis" (UNGP 2012).

2.5 Discussion

Positive deviance has been shown to be effective as a problem-solving approach in certain development domains, but it faces challenges that have so far limited its uptake. Big data could potentially address some of those challenges and/or in other ways enhance current

approaches to positive deviance, providing there was an adapted PD framework that could guide its use. Conversely, positive deviance appears to provide an approach to the detection of socio-economic anomalies that might broaden the application of big data in development. Alongside the growing interest in and practice of both positive deviance and big data in development, this creates an opportunity for "big data-based positive deviance" (BDPD). In this section we will examine how BD could address some of the aforementioned PD challenges and we then propose a means to operationalize their combined use.

2.5.1 BD as a Response to PD Challenges

While big data cannot address all of the positive deviance-related challenges identified in Section 2.3.5, it has potential in relation to most of them:

Time and cost: PD studies mainly use primary data collection both for identification of positive deviants and for PDI: the inquiry into what causes the deviant outcomes. As noted above, primary data collection involves significant time and cost and risk, and use of other forms of data collection could therefore offer important advantages. In light of this, a few studies have made use of traditional secondary data – such as that from surveys – but this brought its own challenges; for example it is hard to identify positive deviants from such datasets as they are often anonymized or out-of-date by the time they become publicly accessible; or it may be difficult to explain causes of positive deviance as important factors are missing from the survey. In addition, while the cost of re-use of survey data may be low, the actual financial costs of the original data-gathering are very high – particularly for census data.

In comparison, big data brings with it not just the gains of reduced time and cost common to re-use of secondary datasets but the more foundational reduction that the costs of initially gathering big data tend to be very low since it often makes use of already existing "data exhaust" from digital processes. In part thanks to low cost, there are also – thinking of satellite imaging and social media data – increasing sources of real-time big data. These avoid the problem of time lag (something particularly challenging with, say, census data which is often only gathered every ten years). Finally, the lack of cost constraints means that big datasets often have a much greater geographical scale than other forms of secondary data.

While big data does still suffer the secondary data shortcoming that it has been created for purposes other than PD analysis, it is increasingly present in locations – such as poorer countries or communities – where survey data either tends to be lacking, or based on very small samples, or inaccurate; these problems themselves sometimes deriving from the high cost and time requirements of surveys and the lack of resources for these locations (Letouzé 2014; Mügge 2014). Initiatives have already demonstrated the ability of big data – such as satellite imaging or CDRs – to fill these data gaps and act as proxies for socio-economic indicators (Njuguna & McSharry 2017; Jean et al. 2016; Henderson et al. 2012).

Big data therefore shows significant potential to help address the time, cost and risk constraints faced by current positive deviance studies; including the constraints associated with using traditional secondary data sources for PD analysis.

Positive deviant identification: As mentioned earlier, the use of primary data collection to identify positive deviants has three main drawbacks: low sample power, inability to identify dynamic anomalies, and limited aggregation analysis. Use of big data sources offers a potential to overcome those challenges as follows:

1a) Sample power: Big data is produced passively at marginal or no additional cost whereas traditional data sources are produced actively with a cost that is proportional to the size of the sample. As a result, BD sources tend to have a much larger coverage of populations. Given that positive deviants are relatively rare, the larger samples from big datasets will enable the identification of a larger number of PDs. Accordingly, the risk of overlooking important factors will be reduced and the ability to generalize practices to larger populations will be improved.

1b) New identification techniques: Although not recognized in the PD literature as a challenge, there is an opportunity offered via BDPD that is not available to traditional positive deviance identification. This is the application of machine learning which, as noted above, works most effectively with large datasets (Hilbert 2016). Machine learning-based approaches for anomaly detection outperform simple statistical models for various types of anomaly (Chandola, Banerjee & Kumar 2009; Goldstein & Uchida 2016), thus providing the potential for better PD identification than currently possible. Supervised machine learning can also be

used for predictive analysis, by first being trained to analyse a small sample set of big data in tandem with ground truth data: for example, satellite images combined with survey data. The survey data already identifies the positive deviants, and machine learning then develops the ability to identify PDs within the corresponding satellite image data. The analytical algorithms can then be applied solely to big data sources and will identify positive deviants from those sources on a much wider scale.

2) Dynamic anomalies: Where traditional data sources typically provide a static, cross-sectional view of behaviour, big data can often provide a dynamic picture of the targeted population over time. Hence, as discussed in Section 2.3.5, BDPD can be better at identifying and potentially eliminating contextual, conditional anomalies as explanations for positive deviance.

3) New levels of aggregation: The majority of PD studies have individuals as the primary unit of analysis, whereas in BD studies, the unit of analysis can range from individuals to communities and regions. Hence, use of big data could provide PD with the ability to identify deviance at different – i.e. higher – levels of aggregation than just individuals.

Methodological risk: Big data is in almost all cases gathered within the explicit intervention of, or tangible visibility to, the subject populations. As such, risks arising from populations knowing they are being observed and questioned, such as the Hawthorne effect or refusal to share practices, are avoided. In addition, where BD sources incorporate outcome indicators of the positive deviant behaviours, then those sources can be used for ongoing monitoring of the effects of a PD intervention. This might help address the challenge that the lack of credible monitoring and evaluation techniques limits uptake of the PD approach, especially in new contexts and settings.

Scalability: As discussed earlier, PD faces two scalability challenges: scaling practices within a community, and scaling practices across communities. Both of these issues are partly rooted in socio-behavioural factors that big data is unlikely to be able to address. But one can hypothesize some potential added value.

For example, for the first challenge, unsupervised machine learning could be applied to a big dataset to cluster the PD intervention population (based on machine-inferred similarities). Then intervention could target only those clusters with socio-economic determinants similar to those of the deviants for practice dissemination and adoption. This could reduce both the time and cost required for scale-up in comparison to the traditional methods of practice dissemination that rely heavily on community mobilization. (Though noting two potential limitations: first that, by definition, positive deviants are sought within populations with similar socio-economic determinants; and second that mobilization and incentives are always likely to be important in any type of PD implementation.) There is also a small possibility that BD-based, statistically-verified evidence might prove more convincing to potential adopters of PD behaviours than the more qualitative findings typical of traditional PD.

The second challenge could be mitigated if cross-community big data sources are available. These would enable the identification of PDs and their behaviour on a broader scale making generalizations possible.

Narrow domain/geographic scope: Despite the effectiveness of PD as a problem-solving approach for international development, its uptake by developing countries in domains other than public health has been very limited, and its application has been concentrated in rural areas of just a few countries. While geographic coverage of BD4D in the literature to date has also been concentrated, that literature already illustrates more application in urban areas and application in several other domains. Big data can thus expand the scope of PD, enabling it to break from its current path dependency. Outline domain examples where big data-based positive deviance could operate include the following:

– Infectious disease control: A number of studies (Tatem et al. 2014; Wesolowski et al. 2014; Wesolowski et al. 2015; Nieto-Sanchez et al. 2015) used CDRs to map the travel patterns of individuals who are members of disease "sources" (areas with many reported disease cases), in order to identify areas vulnerable to transmission, known as "sinks" (areas having high inflows of individuals from source areas). The aim of such studies was to prioritize source and sink areas for disease control. PDs in this case would be "sink" areas with very high potential for transmission due to high travel inflows from source areas, yet having only a small number of reported cases. *Understanding and identifying the measures*

and the factors that were behind this deviance could provide valuable insights into successful disease control for other infected areas.

- Urban resilience/planning: In Zhang et al. (2016), factors affecting waterlogging in one city were used to predict waterlogging in another city using satellite imagery, precipitation meteorological data, terrain data and road maps. *Positive deviance could be used to investigate why certain areas (PDs) within the same city experience less frequent waterlogging than others, and using those factors (e.g. infrastructure, road networks) for better urban planning.*
- Academic research: In Surjandari et al. (2015), Indonesian scholars' publications indexed in Scopus were analysed to map their primary research themes and advise on a nationwide research roadmap. *One could aggregate those publications by department, and identify departments with exceptionally high research publication quality (PDs) as proxied by citation indicators in Scopus or equivalent sources, e.g. using the average h-index*[13] *of the publishing authors. Understanding factors leading to better publication quality would provide insights into departmental-level good practices that could be adopted in other departments.*
- Deforestation: In Burgess et al. (2012), satellite images were used to identify forested districts in Indonesia that are practising illegal logging. *Positive deviance could be applied to identify districts with minimal illegal logging activities (PDs) and then investigate the measures and practices in place within those districts that are linked to that minimization.*
- Agriculture: In Tesfaye et al. (2016), crop, soil and climate data were used to assess the performance of new drought-tolerant crop varieties. *This data could be used to identify those smallholder farms with high productivity (PDs) i.e. high output levels from drought-tolerant crops. Using these or other big datasets, or survey data, one could infer good practices that could be adopted by neighbouring farms facing the same social, economic and environmental constraints.*

2.5.2 PD as an Opportunity for BD4D Applications

[13] Where h is the highest number where a scholar has h papers that have been cited at least h times.

The primary interest of this paper is that outlined in the previous section: to identify current challenges to positive deviance action research, and to identify ways in which big data-based positive deviance might address those challenges. But we can also reverse the polarity of the investigation, and ask what positive deviance might offer the sub-field of big data for development. Reviewing Section 2.4.7, there is little or nothing that positive deviance can do to address issues of representational validity, human capacity, data accessibility, or privacy and legal issues. Instead, these represent constraints that BDPD would have to contend with.

But those working on big data do identify a "need to develop methodologies to characterize and detect socioeconomic anomalies in context" (UNGP 2012). Use of positive deviance to analyse big datasets and to detect anomalies in context cannot be said to provide a theoretical foundation in an academic sense, but it does offer a conceptual frame and a systematic methodology – a theory of change – that can link big data to development outcomes; something which has typically been missing to date. And, at least if all five steps of the positive deviance approach were undertaken, then BDPD helps big data for development move beyond just proof of concept, by creating a real-world impact from the analysis of big data.

2.5.3 Towards Big Data-Based Positive Deviance Analysis

In summary, we have a two-sided argument in favour of a big data-based positive deviance approach. Particularly, using big data instead of – or in conjunction with – traditional primary data sources can potentially address many of the challenges currently faced by positive deviance: reducing time, cost and effort; identifying positive deviants in new or better ways; and enabling positive deviance to break out of its current path dependencies. And, conversely, positive deviance provides a systematic basis for extracting real-world development impacts from big data by putting knowledge about anomalies into action.

We can summarize big data-based positive deviance as follows:

The BDPD approach is a problem-solving asset-based approach that uses big data sources to identify objects (positive deviants) performing unexpectedly well in a specific outcome measure that is digitally recorded, mediated or observed. The primary objective of the BDPD

approach is to identify the behaviours, strategies and factors employed by the positive deviants and develop interventions to facilitate the dissemination and adoption of those strategies.

BDPD objects – the positive deviants – could be individuals, communities, entities, areas, or countries whose uncommon behaviours and strategies, in a specific context, can be translated into a performance measure that is digitally recorded, mediated or observed.

We end with Table 4, which compares the PD and BDPD approaches in terms of the data sources used, the type of anomalies detected, the possible units of analysis, and the employed research methods and techniques.

	Data Sources Used	**Type of Positive Deviants**	**Unit of Analysis**	**Research Methods**	**Data Analysis Methods**
PD	Surveys, Focus Groups, Interviews, Observations	Point Anomalies	Individuals and Entities	Qualitative, Quantitative or Mixed	Statistical Methods
BDPD	Government Data, Online incl. Social Media Data, Mobile Data, Physical Sensor incl. Remote Sensing Data, Offline & Online Surveys, Focus Groups, Interviews, Observations	Point and Contextual (time & spatial) Anomalies	Individuals, Entities, Communities, Regions, Countries, etc.	Quantitative or Mixed	Advanced Analytics and Statistical Methods

Table 4: Comparison between the PD and the BDPD approach

We will be taking forward work applying BDPD, and we hope that other development researchers and practitioners may be encouraged to do the same.

References

Aday, J., Hyden, A., Osking, J. & Tomedi, A. (2016) Hygiene, sanitation, and behaviors that produce positive deviant outcomes in childhood growth in rural eastern Kenya: a qualitative positive deviant investigation, *Annals of Global Health*, 82(3), 437.

Ahrari, M. et al. (2002) Factors associated with successful pregnancy outcomes in upper Egypt: a positive deviance inquiry, *Food and Nutrition Bulletin*, 23(1), 83–88.

Andrews, M. (2015) Explaining positive deviance in public sector reforms in development, *World Development*, 74, 197–208.

Anino, O. C., Were, G. M. & Khamasi, J. W. (2015) Impact evaluation of positive deviance hearth in Migori County, Kenya, *African Journal of Food, Agriculture, Nutrition and Development*, 15(5), 10578–10596.

Aruna, M., Vazir, S. & Vidyasagar, P. (2001) Child rearing and positive deviance in the development of preschoolers: a microanalysis, *Indian Pediatrics*, 38(4), 332–339.

Awofeso, N., Irwin, T. & Forrest, G. (2008) Using positive deviance techniques to improve smoking cessation outcomes in New South Wales prison settings, *Health Promotion Journal of Australia*, 19(1), 72–73.

Babalola, S. (2007) Motivation for late sexual debut in Cote d'Ivoire and Burkina Faso: a positive deviance inquiry, *Journal of HIV/AIDS Prevention in Children and Youth*, 7(2), 65–87.

Babalola, S., Awasum, D. & Quenum-Renaud, B. (2002) The correlates of safe sex practices among Rwandan youth: a positive deviance approach, *African Journal of AIDS Research*, 1(1), 11–21.

Bamberger, M. (2016) Integrating Big Data into the Monitoring and Evaluation of Development Programmes. New York, NY: UN Global Pulse.

Batty, M. et al. (2012) Smart cities of the future, *European Physical Journal: Special Topics*, 214(1), 481–518.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R. & von Schreeb, J. (2011) Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in haiti, *PLoS Medicine*, 8(8), 1–9.

Birhanu, M. Y., Girma, A. & Puskur, R. (2017) Determinants of success and intensity of livestock feed technologies use in Ethiopia: evidence from a positive deviance perspective, *Technological Forecasting and Social Change*, 115, 15–25.

Blumenstock, J. E., Cadamuro, G. & On, R. (2015) Predicting poverty and wealth from mobile phone metadata, *Science*, 350(6264), 1073–1076.

Blumenstock, J. E. & Eagle, N. (2012) Divided we call: disparities in access and use of mobile phones in Rwanda, *Information Technologies & International Development*, 8(2), 1–16.

Bolles, K., Speraw, C., Berggren, G. & Lafontant, J. G. (2002) Ti foyer (hearth) community-based nutrition activities informed by the positive deviance approach in Leogane, Haiti: a programmatic description, *Food and Nutrition Bulletin*, 23(4 Supp), 11–17.

Bradley, E. H., Curry, L. A., Ramanadhan, S., Rowe, L., Nembhard, I. M. & Krumholz, H. M. (2009) Research in action: using positive deviance to improve quality of health care, *Implementation Science*, 4, 25.

Buckee, C. O., Wesolowski, A., Eagle, N. N., Hansen, E. & Snow, R. W. (2013) Mobile phones and malaria: modeling human and parasite travel, *Travel Medicine and Infectious Disease*, 11(1), 15–22.

Burgess, R., Hansen, M., Olken, B. A., Potapov, P. & Sieber, S. (2012) The political economy of deforestation in the tropics, *The Quarterly Journal of Economics*, 127(4), 1707–1754.

Cartesian (2014) *Using Mobile Data for Development*. Boston, MA: Cartesian.

Chandola, V., Banerjee, A. & Kumar, V. (2009) Anomaly detection, *ACM Computing Surveys*, 41(3), 1–58.

Chandy, R., Hassan, M. & Mukherji, P. (2017) Big data for good: insights from emerging markets, *Journal of Product Innovation Management*, 34(5), 703–713.

D'Alimonte, M. R., Deshmukh, D., Jayaraman, A., Chanani, S. & Humphries, D. L. (2016) Using positive deviance to understand the uptake of optimal infant and young child feeding practices by mothers in an urban slum of Mumbai, *Maternal and Child Health Journal*, 20(6), 1133–1142.

Delias, P. (2017) A positive deviance approach to eliminate wastes in business processes, *Industrial Management & Data Systems*, 117(7), 1323–1339.

Doll, C. N. H. & Pachauri, S. (2010) Estimating rural populations without access to electricity in developing countries through night-time light satellite imagery, *Energy Policy*, 38(10), 5661–5670.

Escobar, N. M. O. et al. (2017) Using positive deviance in the prevention and control of MRSA infections in a Colombian hospital: a time-series analysis, *Epidemiology and Infection*, 145(5), 981–989.

Felt, L. J. (2011) Present Promise, Future Potential: Positive Deviance and Complementary Theory. Unpublished manuscript. http://www.laurelfelt.org/wp-content/uploads/2011/06/PositiveDeviance-CodyMayer.LaurelFelt.Quals_.May2011.pdf

Frias-Martinez, V. & Virseda, J. (2013) Cell phone analytics: scaling human behavior studies into the millions, *Information Technologies & International Development*, 9(2), 35–50.

Garrett, J. J. & Barrington, C. (2013) 'We do the impossible': women overcoming barriers to cervical cancer screening in rural Honduras – a positive deviance analysis, *Culture, Health & Sexuality*, 15(6), 637–651.

Gartner (2013) *Big Data*. Stamford, CT: Gartner. https://www.gartner.com/it-glossary/big-data

Gidado, M., Obasanya, J. O., Adesigbe, C., Huji, J. & Tahir, D. (2010) Role of positive deviants among leprosy self-care groups in leprosy settlement, Zaria, Nigeria, *Journal of Community Medicine and Primary Health Care*, 22(1–2).

Goldstein, M. & Uchida, S. (2016) A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data, *PLoS ONE*, 11(4).

Graham, M. & Shelton, T. (2013) Geography and the future of big data, big data and the future of geography, *Dialogues in Human Geography*, 3(3), 255–261.

Greenhalgh, T. & Peacock, R. (2005) Effectiveness and efficiency of search methods in systematic reviews of complex evidence: audit of primary sources, *British Medical Journal*, 331(7524), 1064–1065.

Guldan, G. S. et al. (1993) Weaning practices and growth in rural sichuan infants: a positive deviance study, *Journal of Tropical Pediatrics*, 39(3), 168–175.

Haas, E. M., Bartholomé, E. & Combal, B. (2009) Time series analysis of optical remote sensing data for the mapping of temporary surface water bodies in sub-Saharan western Africa, *Journal of Hydrology*, 370(1–4), 52–63.

Henderson, J. V., Storeygard, A. & Weil, D. N. (2012) Measuring economic growth from outer space, *American Economic Review*, 102(2), 994–1028.

Hendrickson, J. L., Dearden, K., Pachón, H., An, N. H., Schroeder, D. G. & Marsh, D. R. (2002) Empowerment in rural Viet Nam: exploring changes in mothers and health volunteers in the context of an integrated nutrition project, *Food and Nutrition Bulletin*, 23(4 Supp), 86–94.

Herington, M. J. & van de Fliert, E. (2018) Positive deviance in theory and practice: a conceptual review, *Deviant Behavior*, 39(5), 664–678.

Hilbert, M. (2016) Big data for development: a review of promises and challenges, *Development Policy Review*, 34(1), 135–174.

Huang, Z., Cao, F., Jin, C., Yu, Z. & Huang, R. (2017) Carbon emission flow from self-driving tours and its spatial relationship with scenic spots – a traffic-related big data method, *Journal of Cleaner Production*, 142, 946–955.

Ifaei, P., Karbassi, A., Lee, S. & Yoo, C. (2017) A renewable energies-assisted sustainable development plan for Iran using techno-econo-socio-environmental multivariate analysis and big data, *Energy Conversion and Management*, 153(October), 257–277.

Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B. & Ermon, S. (2016) Combining satellite imagery and machine learning to predict poverty, *Science*, 353(6301), 790–794.

Jiang, W., Jia, K., Chen, Z., Deng, Y. & Rao, P. (2017) Using spatiotemporal remote sensing data to assess the status and effectiveness of the underground coal fire suppression efforts during 2000–2015 in Wuda, China, *Journal of Cleaner Production*, 142, 565–577.

Kanani, S. & Popat, K. (2012) Growing normally in an urban environment: positive deviance among slum children of Vadodara, India, *Indian Journal of Pediatrics*, 79(5), 606–611.

Khabsa, M. & Giles, C. L. (2014) The number of scholarly documents on the public web, *PLoS ONE*, 9(5).

Kim, Y. M., Heerey, M. & Kols, A. (2008) Factors that enable nurse-patient communication in a family planning context: a positive deviance study, *International Journal of Nursing Studies*, 45(10), 1411–1421.

Kohli, D., Sliuzas, R., Kerle, N. & Stein, A. (2012) An ontology of slums for image-based classification, *Computers, Environment and Urban Systems*, 36(2), 154–163.

Lackovich-Van Gorp, A. (2017) Unearthing local forms of child protection: positive deviance and abduction in Ethiopia, *Action Research*, 15(1), 39–52.

Lapping, K., Marsh, D.R., Rosenbaum, J., Swedberg, E., Sternin, J., Sternin, M. & Schroeder, D.G. (2002) The positive deviance approach: Challenges and opportunities for the future, *Food and Nutrition Bulletin*, 23(4 Supp1), 128–135.

Lapping, K., Schroeder, D., Marsh, D., Albalak, R. & Jabarkhil, M. Z. (2002b) Comparison of a positive deviant inquiry with a case-control study to identify factors associated with nutritional status among Afghan refugee children in Pakistan, *Food and Nutrition Bulletin*, 23(4 Supp2), 26–33.

LeMahieu, P. G., Nordstrum, L. E. & Gale, D. (2017) Positive deviance: learning from positive anomalies, *Quality Assurance in Education*, 25(1), 109–124.

Letouzé, E. & Jütting, J. (2015) *Official Statistics, Big Data and Human Development: Towards a New Conceptual and Operational Approach*. Data Pop Alliance. http://www.paris21.org/sites/default/files/WPS_OfficialStatistics_June2015.pdf

Levinson, F. J., Barney, J., Bassett, L. & Schultink, W. (2007) Utilization of positive deviance analysis in evaluating community-based nutrition programs: an application to the Dular program in Bihar, India, *Food and Nutrition Bulletin*, 28(3), 259–265.

Li, S., Ye, X., Lee, J., Gong, J. & Qin, C. (2017) Spatiotemporal analysis of housing prices in China: a big data perspective, *Applied Spatial Analysis and Policy*, 10(3), 421–433.

Long, K. N. et al. (2013) Determinants of better health: a cross-sectional assessment of positive deviants among women in West Bengal, *BMC Public Health*, 13, 372.

Lu, X. et al. (2016) Unveiling hidden migration and mobility patterns in climate stressed regions: a longitudinal study of six million anonymous mobile phone users in Bangladesh, *Global Environmental Change*, 38, 1–7.

Lu, X., Bengtsson, L. & Holme, P. (2012) Predictability of population displacement after the 2010 Haiti earthquake, *Proceedings of the National Academy of Sciences*, 109(29), 11576–11581.

Luo, X. et al. (2017) Analysis on spatial-temporal features of taxis' emissions from big data informed travel patterns: a case of Shanghai, China, *Journal of Cleaner Production*, 142, 926–935.

de Macedo, R. de C. R. et al. (2012) Positive deviance: using a nurse call system to evaluate hand hygiene practices, *American Journal of Infection Control*, 40(10), 946–950.

Mackintosh, U. A., Marsh, D. R. & Schroeder, D. G. (2002) Sustained positive deviant child care practices and their effects on child growth in Viet Nam, *Food and Nutrition Bulletin*, 23(4 Supp), 18–27.

Marra, A. R. et al. (2010) Positive deviance: a new strategy for improving hand hygiene compliance, *Infection Control & Hospital Epidemiology*, 31(1), 12–20.

Marra, A. R. et al. (2011) Positive deviance: a program for sustained improvement in hand hygiene compliance, *American Journal of Infection Control*, 39(1), 1–5.

Marra, A. R. et al. (2013) A multicenter study using positive deviance for improving hand hygiene compliance, *American Journal of Infection Control*, 41(11), 984–988.

Marsh, D. R. et al. (2002) Identification of model newborn care practices through a positive deviance inquiry to guide behavior-change interventions in Haripur, Pakistan, *Food and Nutrition Bulletin*, 23(4 Supp), 109–18.

Marsh, D. R., Schroeder, D. G., Dearden, K. A., Sternin, J. & Sternin, M. (2004) The power of positive deviance, *British Medical Journal*, 329(7475), 1177–1179.

Mashey, J. R. (1998) Big data and the next wave of infrastress, paper presented at *Computer Systems Laboratory Colloquium*, Stanford University, CA, 25 Feb.

Merchant, S. S. & Udipi, S. A. (1997) Positive and negative deviance in growth of urban slum children in Bombay, *Food and Nutrition Bulletin*, 18(4), 323–336.

Merita, M., Sari, M. T. & Hesty, H. (2017) The positive deviance of feeding practices and carring with nutritional status of toddler among poor families, *Jurnal Kesehatan Masyarakat*, 13(1), 106–112.

Mertens, W., Recker, J., Kohlborn, T. & Kummer, T.-F. (2016) A framework for the study of positive deviance in organizations, *Deviant Behavior*, 37(11), 1288–1307.

Moher, D., Liberati, A., Tetzlaff, J. & Altman, D. G. (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, *British Medical Journal*, 339, b2535.

Mügge, D. (2014) Poor Numbers: How We are Misled by African Development Statistics and What to Do About It, by Morten Jerven, *Review of International Political Economy*, 21(5), 1131–1133.

Ndiaye, M., Siekmans, K., Haddad, S. & Receveur, O. (2009) Impact of a positive deviance approach to improve the effectiveness of an iron-supplementation program to control nutritional anemia among rural Senegalese pregnant women, *Food and Nutrition Bulletin*, 30(2), 128–136.

Nel, H. (2018) A comparison between the asset-oriented and needs-based community development approaches in terms of systems changes, *Practice*, 30(1), 33–52.

Nieto-Sanchez, C., Baus, E. G., Guerrero, D. & Grijalva, M. J. (2015) Positive deviance study to inform a chagas disease control program in southern Ecuador, *Memorias do Instituto Oswaldo Cruz*, 110(3), 299–309.

Nishat, N. & Batool, I. (2011) Effect of 'Positive Hearth Deviance' on feeding practices and underweight prevalence among children aged 6-24 months in Quetta district, Pakistan: a comparative cross sectional study, *Sri Lanka Journal of Child Health*, 40(2), 57–62.

Njuguna, C. & McSharry, P. (2017) Constructing spatiotemporal poverty indices from big data, *Journal of Business Research*, 70, 318–327.

OECD (2017) *DAC List of ODA Recipients*. Paris: Organisation for Economic Co-operation and Development.

Okoli, C. (2015) A guide to conducting a standalone systematic literature review, *Communications of the AIS*, 37, 879–910.

Osborne, J. W. & Overbay, A. (2004) The power of outliers (and why researchers should always check for them), *Practical Assessment, Research & Evaluation*, 9(6), 1–8.

Pascale, R., Sternin, J. & Sternin, M. (2010) The Power of Positive Deviance: How Unlikely Innovators Solve the World's Toughest Problems. Boston, MA: Harvard Business Press.

Petticrew, M. & Roberts, H. (2005) *Systematic Reviews in the Social Sciences: A Practical Guide*. Chichester, UK: John Wiley & Sons.

Pfeffer, K., Verrest, H. & Poorthuis, A. (2015) Big data for better urban life? - an exploratory study of critical urban issues in two Caribbean cities: Paramaribo (Suriname) and Port of Spain (Trinidad and Tobago), *European Journal of Development Research*, 27(4), 505–522.

Positive Deviance Initiative (2010) *Basic Field Guide to the Positive Deviance Approach*. Medford, MA: Positive Deviance Initiative, Tufts University.

Roche, M. L., Marquis, G. S., Gyorkos, T. W., Blouin, B., Sarsoza, J. & Kuhnlein, H. V. (2017) A community-based positive deviance/hearth infant and young child nutrition intervention in ecuador improved diet and reduced underweight, *Journal of Nutrition Education and Behavior*, 49(3), 196–203.

Saïd Business School (2010) *Exploring Positive Deviance – New Frontiers in Collaborative Change*. Oxford, UK: Saïd Business School, University of Oxford.

Sengupta, R., Heeks, R., Chattapadhyay, S. & Foster, C. (2017) *Exploring Big Data for Development: An Electricity Sector Case Study from India*, Development Informatics Working Paper 66. Manchester, UK: Global Development Institute, University of Manchester.

Sethi, V., Kashyap, S., Seth, V. & Agarwal, S. (2003) Encouraging appropriate infant feeding practices in slums: a positive deviance approach, *Pakistan Journal of Nutrition*, 2(3), 164–166.

Sethi, V., Sternin, M., Sharma, D., Bhanot, A. & Mebrahtu, S. (2017) Applying positive deviance for improving compliance to adolescent anemia control program in tribal communities of India, *Food and Nutrition Bulletin*, 38(3), 447–452.

Shekar, M., Habicht, J. P. & Latham, M. C. (1991) Is positive deviance in growth simply the converse of negative deviance?, *Food and Nutrition Bulletin*, 13(1), 7–11.

Shekar, M., Habicht, J. P. & Latham, M. C. (1992) Use of positive-negative deviant analyses to improve programme targeting and services: example from the Tamil Nadu integrated nutrition project, *International Journal of Epidemiology*, 21(4), 707–713.

Shoenberger, N. A. (2017) Bridging normative and reactivist perspectives: an introduction to positive deviance, in *Routledge Handbook on Deviance*, S. E. Brown & O. Sefina (eds). Abingdon, UK: Routledge, 42–51.

Singhal, A. (2011) Turning diffusion of innovation paradigm on its head: the positive deviance approach to social change, in *The Diffusion of Innovations*, A. Vishwanath & G. A. Barnett (eds). New York, NY: Peter Lang, 193–205.

Song, M. & Wang, S. (2017) Participation in global value chain and green technology progress: evidence from big data of Chinese enterprises, *Environmental Science and Pollution Research*, 24(2), 1648–1661.

Springer, A., Nielsen, C. & Johansen, I. (2016) *Positive Deviance by the Numbers*. Medford, MA: Positive Deviance Initiative, Tufts University.

Sternin, J. (2002) Positive deviance: a new paradigm for addressing today's problems today, *The Journal of Corporate Citizenship*, 5(Spring), 57–63.

Sternin, J. & Choo, R. (2000) The power of positive deviancy, *Harvard Business Review*, 78(1), 1–3.

Sternin, M., Sternin, J. & Marsh, D. (1997) Rapid Sustained Childhood Malnutrition Alleviation Through a Positive-Deviance Approach in Rural Vietnam: Preliminary Findings. Arlington, VA: Partnership for Child Health Care, BASICS.

Sternin, M., Sternin, J. & Marsh, D. (1998) *Designing a Community-Based Nutrition Program Using the Hearth Model and the Positive Deviance Approach: A Field Guide.* Westport, CT: Save the Children.

Su, S., Li, Z., Xu, M., Cai, Z. & Weng, M. (2017) A geo-big data approach to intra-urban food deserts: transit-varying accessibility, social inequalities, and implications for urban planning, *Habitat International*, 64, 22–40.

Surjandari, I., Dhini, A., Lumbantobing, E. W. I., Widari, A. T. & Prawiradinata, I. (2015) Big data analysis of indonesian scholars' publications: a research theme mapping, *International Journal of Technology*, 6(4), 650–658.

Sutton, P., Elvidge, C. & Ghosh, T. (2007) Estimation of gross domestic product at sub-national scales using nighttime satellite imagery, *International Journal of Ecological Economics & Statistics*, 8(S07), 5–21.

Tang, H., Yang, X. & Zhang, Y. (2014) Effort at constructing big data sensor networks for monitoring greenhouse gas emission, *International Journal of Distributed Sensor Networks*, 10(7).

Tatem, A. J. et al. (2014) Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning, *Malaria Journal*, 13, 52.

Tatem, A. J., Noor, A. M., von Hagen, C., Di Gregorio, A. & Hay, S. I. (2007) High resolution population maps for low income nations: combining land cover and census in East Africa, *PLoS ONE*, 2(12).

Tatem, A. J., Qiu, Y., Smith, D. L., Sabot, O., Ali, A. S. & Moonen, B. (2009) The use of mobile phone data for the estimation of the travel patterns and imported plasmodium falciparum rates among Zanzibar residents, *Malaria Journal*, 8, 287.

Taylor, L. & Broeders, D. (2015) In the name of development: power, profit and the datafication of the global South, *Geoforum*, 64, 229–237.

Tekle, L. (2015) Analysis of positive deviance farmer training centers in Northern Ethiopia, *American Journal of Rural Development*, 3(1), 10–14.

Tesfaye, K. et al. (2016) Targeting drought-tolerant maize varieties in Southern Africa: a geospatial crop modeling approach using big data, *International Food and Agribusiness Management Review*, 19(A), 75–92.

Tong, T. & Li, H. (2017) Demand for MOOC - an application of big data, *China Economic Review*, in press.

UNGP (2012) *Big Data for Development: Challenges & Opportunities.* New York, NY: UN Global Pulse.

Vossenaar, M. et al. (2009) The positive deviance approach can be used to create culturally appropriate eating guides compatible with reduced cancer risk, *Journal of Nutrition*, 139(4), 755–762.

Vossenaar, M., Bermúdez, O. I., Anderson, A. S. & Solomons, N. W. (2010) Practical limitations to a positive deviance approach for identifying dietary patterns compatible with the reduction of cancer risk, *Journal of Human Nutrition and Dietetics*, 23(4), 382–392.

Wesolowski, A. et al. (2014) Quantifying travel behavior for infectious disease research: a comparison of data from surveys and mobile phones, *Scientific Reports*, 4, 5678.

Wesolowski, A. et al. (2015) Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data, *Proceedings of the National Academy of Sciences*, 112(35), 11114–11119.

Wilson, C. (2011) Digital media in the Egyptian revolution: descriptive analysis from the Tahrir data sets, *International Journal of Communication*, 5, 1248–1272.

Wilson, R. et al. (2016) Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal earthquake, *PLoS Currents Disasters*, 24 Feb, 8.

Wishik, S. M. & Van Der Vynckt, S. (1976) The use of nutritional "positive deviants" to identify approaches for modification of dietary practices, *American Journal of Public Health*, 66(1), 38–42.

Wollinka, O., Keeley, E., Burkhalter, B. R. & Bashir, N. (1997) *Hearth Nutrition Model: Applications in Haiti, Viet Nam and Bangladesh*. Arlington, VA: Partnership for Child Health Care, BASICS.

WP (2018) *Open Data*. Wikipedia. https://en.wikipedia.org/wiki/Open_data

Zaidi, Z., Jaffery, T., Shahid, A., Moin, S., Gilani, A. & Burdick, W. (2012) Change in action: using positive deviance to improve student clinical performance, *Advances in Health Sciences Education*, 17(1), 95–105.

Zeitlin, M. (1991) Nutritional resilience in a hostile environment: positive deviance in child nutrition, *Nutrition Reviews*, 49(9), 259–268.

Zeitlin, M. F. et al. (1990) Positive Deviance in Child Nutrition: With Emphasis on Psychosocial and Behavioural Aspects and Amplications for Development. Tokyo: United Nations University.

Zeitlin, M., Ghassemi, H. & Mansour, M. (1994) Positive deviance in child nutrition: a discussion, *Ecology of Food and Nutrition*, 31(3–4), 295–302.

Zhang, N., Chen, H., Chen, J. & Chen, X. (2016) Social media meets big urban data: a case study of urban waterlogging analysis, *Computational Intelligence and Neuroscience*, 2016.

Chapter Three: Publication Outperformance among Global South Researchers: An Analysis of Individual-Level and Publication-Level Predictors of Positive Deviance

Basma Albanna, Richard Heeks and Julia Handl

Abstract

Research and development are central to economic growth, and a key challenge for countries of the global South is that their research performance lags behind that of the global North. Yet, among Southern researchers, a few significantly outperform their peers and can be styled research "positive deviants" (PDs). In this paper we ask: who are those PDs, what are their characteristics and how are they able to overcome some of the challenges facing researchers in the global South? We examined a sample of 203 information systems researchers in Egypt who were classified into PDs and non-PDs (NPDs) through an analysis of their publication and citation data. Based on six citation metrics, we were able to identify and group 26 PDs. We then analysed their attributes, attitudes, practices, and publications using a mixed-methods approach involving interviews, a survey and analysis of publication-related datasets. Two predictive models were developed using partial least squares regression; the first predicted if a researcher is a PD or not using individual-level predictors and the second predicted if a paper is a paper of a PD or not using publication-level predictors. PDs represented 13% of the researchers but produced about half of all publications, and had almost double the citations of the overall NPD group. At the individual level, there were significant differences between both groups with regard to research collaborations, capacity development, and research directions. At the publication level, there were differences relating to the topics pursued, publication outlets targeted, and paper features such as length of abstract and number of authors.

3.1 Introduction

A nation's scientific research capability, characterised by its direct engagement in the creation of knowledge, plays a vital role in its sustainable economic development, and the strong correlation between science and technology development and economic development is well documented (King 2004; Man et al. 2004). Scientific research is required both to create the new technologies and techniques that increase local productivity and economic growth, and to adapt technologies imported from abroad (Goldemberg 1998). A necessary part of this, in order to build a strong knowledge society with a thriving 'culture of science', is the publication and dissemination of research results (Salager-Meyer 2008)[14].

A clear research divide is visible between the global South[15] and the global North. This can be seen in terms of research investment and capability. For example, the average national expenditure on research and development from 2005-2014 was 1.44% of GDP in Northern countries but only 0.38% of GDP in Southern countries (Blicharska et al. 2017) while the number of researchers per million population in 2017 was 4,351 in the global North and 713 in the global South (World Bank Group 2020). The divide is also manifest in scientific outputs. In 2018, global North countries produced an average of more than 35,000 scientific and technical journal articles per country while global South countries produced an average of 9,700, or 4,000 if China and India are excluded[16] (World Bank Group 2020). Despite some signs of progress, there also remains an important gap in terms of per-country and per-researcher citation rates between North and South (Gonzalez-Brambila et al. 2016, Confraria et al. 2017). The divide in terms of highly-cited outputs is even starker, with global South

14 Acknowledging that this takes a Western perspective on knowledge; a perspective that has been critiqued given the other forms of non-Western knowledge and knowledge production that exist (Thesee 2006)

15 The terms "South" and "Southern" will be used to refer to countries classified as upper-middle income, lower-middle income, and low income. Accordingly, the terms "North" and "Northern" will be used to refer to countries that are members of the OECD (Organisation for Economic Co-operation and Development) or are classified as high-income economies by the World Bank based on estimates of gross national income per capita.

16 Small island states and non-UN-member territories are excluded from this calculation.

researchers (again excluding China and India) authoring less than 2% of the top 1% most-cited articles globally (National Science Board 2018).

It is this latter issue – low citation rates for Southern research – that forms our particular focus in this paper, and for which a number of explanations have been put forward. Statistical evidence shows that the lower levels of investment and lower relative populations of researchers in the global South are key factors; the latter issue is exacerbated by the brain drain of Southern researchers who relocate to the global North (Man et al. 2004; Salager-Meyer 2008; Pasgaard & Strange 2013). Lower levels of English language proficiency are also a factor, given the skew of international journal publication towards English (Man et al. 2004; Gonzalez-Brambila et al. 2016, Confraria et al. 2017). Other recognised institutional exclusion factors and / or biases against Southern researchers include difficulty in securing research grants (Karlsson et al. 2007), and a greater likelihood that reviewers and editors of mainstream scientific journals will reject a paper from a global South institution than a paper of equivalent quality from a global North institution (Gibbs 1995; Leimu & Koricheva 2005).

Among the valuable research conducted on this issue to date, there have been three main approaches: country-level statistical analysis, paper-level statistical analysis, or individual-level analysis. While the latter includes author-related factors, Southern researchers as individuals are rarely investigated. In particular, there has been no previous research focusing on "exceptions to the rule": those few Southern researchers who are able to achieve much higher research performance than their peers. *Pre hoc*, it is reasonable to hypothesise that such researchers could provide valuable insights and lessons that might help to better understand and even mitigate the current North-South divide in research outputs and citation. It is therefore the purpose of this paper to specifically study these exceptions by investigating what characterises high-performing researchers and their publications in a global South context.

In order to do this, we make use of the "positive deviance" (PD) approach, given that this attempts to systematically identify and learn from "outliers" – individuals who are performing substantially better than expected and better than their peers, given the resources and socio-economic conditions they are exposed to (Sternin et al. 1997). First used at scale in order to learn from Vietnamese families with well-nourished children in contexts of widespread

malnutrition (Sternin 2002), the positive deviance approach has subsequently spread to other domains (Albanna & Heeks 2019). However, the conventional PD approach relies heavily on primary data collection to develop a baseline from which positive deviants (PDs) are identified — a process that is both time- and labour-intensive with costs directly proportional to sample size. Recent developments in the availability of digital datasets have presented possibilities for identification and understanding of PDs in new and better ways (Albanna & Heeks 2019). This new digital data-powered approach to positive deviance was seen as particularly relevant for investigation of scientific researchers given the existence of platforms that digitally index and/or analyse the scholarly work of researchers, enabling evaluation of their performance through multiple dimensions and metrics.

For this study, we chose a sample of 203 information systems (IS) researchers from Egypt to identify factors that enabled a few positive deviants to outperform their peers. Positive deviants are defined as researchers who outperform their peers in both productivity (articles published) and impact (article citations). They were identified based on six citation metrics that take into account those two dimensions of performance in different ways. We conducted an analysis of the researchers' attributes, attitudes, practices, and publications, based on a mixed-methods approach that employed interviews, surveys, and analysis of publication-related datasets. Two methodological innovations were developed in this study. The first was the use of multiple performance metrics in identifying PDs, which enabled us to profile PDs into groups based on those metrics. The second was the identification of extrinsic and intrinsic predictors of PDs' publications as a way of understanding and reflecting on some of their publication strategies. Hence, this paper has two main contributions. The first is *contextual,* which is the identification of predictors of high performers or PDs in a Southern country, who face challenges different from those facing researchers in Northern countries. And the second is *methodological*, where a combination of multiple performance indicators and a number of data sources were used to develop a holistic approach for identifying, profiling and characterising PDs.

In what follows, we first present a review of related work on high-performing researchers before explaining the data sources and methodology of this positive deviance approach. The methodology steps are then undertaken: defining the study focus, determining the positive

deviants, and discovering the features of positive deviants and their published papers. We end with discussion and conclusions.

3.2 Related Work

There is a substantial body of research on the predictors of individual-level high research performance over the last four decades. While the terminology of "positive deviants" has not been used; analogous concepts and synonymous terms have been. Relevant literature on *highly productive academics* includes work studying their attitudes, practices and perceptions in 11 European countries (Kwiek 2016), and their attributes, perceptions and structural predictors in China, Japan and South Korea (Postiglione & Jung 2013); a series of studies that investigated the characteristics and work habits of the top (three or four) educational psychology researchers in the US (Kiewra & Creswell 2000; Patterson-Hazley & Kiewra 2013) and in Germany (Flanigan et al. 2018); and a paper on the strategies and attributes of highly productive academics in school psychology, who were mainly Americans (Martínez et al. 2011). Other research that has looked into *top performers* includes Kwiek's study (2018), which investigated both individual and institutional variables to identify predictors of research success for the top 10% of Polish academics, and Kelchtermans & Veugelers' (2013) study on top performing Belgian researchers, which investigated the effects of co-authorship, gender and previous top performance. There are also studies on *research stars* such as the study by Yair et al. (2017) on Israeli Prize laureates in life and exact sciences; and the study by White et al. (2012) on American researchers in business schools, where individual and situational variables were explored. *High achievers* were identified in a study by Harris & Kaine (1994) that investigated the preferences and perceptions of high-performing Australian university economists and *highly cited scientists* were studied by Parker et al. (2010) and Parker et al. (2013) who sought to identify the social characteristics and opinions of the 0.1% most cited environmental scientists and ecologists worldwide. *Eminent scientists* were studied by Prpić (1996) to explore the most important predictors of productivity among Croatian scientists and *top producing* researchers were identified in a study by Mayrath (2008) which aimed at understanding the attributes of the authors having the most publications in educational psychology journals.

Beneath the factors identified in these studies have lain theoretical models proposed to explain the research performance and outperformance observed, of which three will be mentioned here. The *sacred spark theory* (Cole & Cole 1973) states that highly productive researchers have an inner drive and motivation to do science that is fuelled by their love of the work. Other theories look more at the external environment. *Utility maximisation theory* (Kyvik 1990) argues that the extent to which researchers research and publish – as opposed to other activities – is determined by the personal utility or benefit they perceive themselves getting; that utility often being significantly determined by external incentives or disincentives that attach to the different activities. *Cumulative advantage theory* (Merton 1968) is somewhat similar in identifying external reward systems, and their reinforcement or otherwise of research and publication activity, as important (Cole & Cole 1973). But it sees researchers who begin with some advantage (either innate or external) being increasingly more productive over time compared to others as they gain further advantage, such as greater likelihood of obtaining research grants, or participating in collaborations.

As summarised next, this work has been of significant value in providing insights into high-performing researchers. However, we can also identify three lacunae which the current paper seeks to address.

First, geographic. It is evident from the above that there is a geographic concentration of such studies on high-income countries of the global North: the one study including China is the sole exception. There has thus been practically no consideration of research performance in the resource-constrained countries of the global South. Addressing this unexplored topic is particularly pressing, given the imperative to improve the contribution of research to national development in these countries, and given that findings from such a study might lead to new context-aware predictors of high research performance that could mitigate some of the challenges reflected in the current North-South divide. Hence the justification for the current paper.

Second, methodological in relation to the dependent variable of performance measurement. The majority of studies identify and rank high-performing researchers based on (i) productivity, as measured by number of articles published (e.g. Harris & Kaine 1994; Prpić 1996; Postiglione & Jung 2013; Kwiek 2016) or, in case of some studies, (ii) impact as measured

by number of citations (e.g. Parker et al. 2010; Parker et al. 2013). There are clearly benefits in incorporating both productivity and impact measures yet this was rarely found in the literature reviewed (Altanopoulou et al. 2012). Recent advances in citation metrics and availability of tools such as Harzing's Publish or Perish software (Harzing 2007) provide an opportunity to measure performance along different dimensions[17] and using combined measures. The current study therefore combines a number of citation metrics to evaluate researchers; enabling a balanced consideration of both productivity and impact and allowing control for factors like article and author age.

Third, methodological in relation to the independent variables or predictors. Table 5 presents the significant predictors of high performers in research, identified from previous studies and forming a foundation for modelling and analysis for the current study. These can be grouped into seven main categories: *Personal or Demographic* characteristics such as age, gender and education; *Internationalisation and Research Collaboration* such as participation in domestic or international research teams; *Research Engagement* with publishing entities; *Research Approach* including focus*; Academic Roles* covering distribution of time between different academic activities; *Practices* associated with undertaking research; and *Institution* predictors related to work environment actuality or preferences.

Significant Predictors	**References**
Personal/Demographics	
Gender: being male	(Prpić 1996) (Parker et al. 2010) (Patterson-Hazley & Kiewra 2013) (Kwiek 2016)
Being older in age or in years of active publication	(Prpić 1996) (Parker et al. 2010) (Patterson-Hazley & Kiewra 2013) (Kwiek 2016)
Younger age of obtaining PhD	(Prpić 1996)
Holding academic rank of professor	(Kelchtermans & Veugelers 2013) (Kwiek 2016)
Training in top programmes in top schools	(Kiewra & Creswell 2000)
Having outside interests	(Kiewra & Creswell 2000)
Having notable figures as advisors	(Mayrath 2008) (Flanigan et al. 2018)

[17] https://harzing.com/pophelp/metrics.htm

Ability in more than two foreign languages	(Prpić 1996)
Good writing skills	(Kiewra & Creswell 2000) (Mayrath 2008)
Passion, curiosity and/or deep interest in research	(Mayrath 2008)
Internationalisation and Research Collaborations	
Connectivity with top researchers	(Patterson-Hazley & Kiewra 2013)
Domestic research collaborations	(Harris & Kaine 1994) (Mayrath 2008) (Kwiek 2016)
International research collaborations	(Harris & Kaine 1994) (Prpić 1996) (Postiglione & Jung 2013) (Kwiek 2016) (Kwiek 2018)
Publishing abroad	(Harris & Kaine 1994) (Prpić 1996) (Kwiek 2016) (Kwiek 2018)
Research is international in scope	(Kwiek 2016) (Kwiek 2018)
Co-authorship: more co-authors	(Prpić 1996) (Kelchtermans & Veugelers 2013)
Research Engagement	
Engaging in peer review	(Prpić 1996) (Kiewra & Creswell 2000) (Kwiek 2016)
Being editor of journals/book series	(Harris & Kaine 1994) (Prpić 1996) (Kiewra & Creswell 2000) (Kwiek 2016) (Kwiek 2018)
Research Approach	
Focus on basic/theoretical research	(Parker et al. 2010) (Postiglione & Jung 2013) (Kwiek 2016)
Focus on pioneering science i.e. exploring novel, under-researched issues	(Kiewra & Creswell 2000)
Constant research focus i.e. finding a niche and carving it out	(Kiewra & Creswell 2000) (Mayrath 2008)
Research leading to acceptable results and not necessarily spectacular results	(Harris & Kaine 1994)
Research which looks for immediate solutions	(Harris & Kaine 1994)
Research that will enhance reputation and prospects for promotion	(Harris & Kaine 1994)
Academic Roles	
Having an administrative role	(Patterson-Hazley & Kiewra 2013) (Kelchtermans & Veugelers 2013)
More time available for research: lower teaching load	(White et al. 2012) (Postiglione & Jung 2013) (Kwiek 2016) (Kwiek 2018)
Mentoring/supervising students	(Prpić 1996) (Kiewra & Creswell 2000) (Mayrath 2008) (White et al. 2012)

Practices	
Time management practices/skills (e.g. stability in daily routines and working for long hours, fixed academic writing time)	(Kiewra & Creswell 2000) (Mayrath 2008) (Parker et al. 2010) (White et al. 2012) (Kwiek 2016) (Flanigan et al. 2018)
Research management strategies (e.g. meeting weekly with collaborators)	(Mayrath 2008) (Flanigan et al. 2018)
Publishing professional and/or scientific works during their undergraduate studies	(Prpić 1996)
Receiving feedback on manuscripts from colleagues or mentors	(Mayrath 2008)
Very good knowledge of the literature	(Mayrath 2008)
Setting deadlines	(White et al. 2012)
Institution	
Research is considered in HR decisions such as promotion	(Kwiek 2016)
Workplace has a strong performance orientation	(Postiglione & Jung 2013) (Kwiek 2016)
Workplace perceived as conducive to research	(Harris & Kaine 1994)
Workplace perceived as a relaxed environment	(Harris & Kaine 1994)
Workplace perceived as an environment that provides opportunity to work on challenging problems	(Harris & Kaine 1994)

Table 5: Significant predictors of high-performing researchers from previous studies

These predictors were almost all identified using qualitative methods such as interviews (Kiewra & Creswell 2000; Flanigan et al. 2018), quantitative methods such as surveys (Harris & Kaine 1994; Prpić 1996; Kwiek 2016; Kwiek 2018; Postiglione & Jung 2013) or a mix of both (Mayrath 2008; Martínez et al. 2011; Patterson-Hazley & Kiewra 2013). What is consistent here is the focus on individual researcher-level data. Largely missing has been publication-level data.

Yet there are a number of reasons for thinking that adding in publication-level data can provide valuable additional insights. Publication data can provide insights into some of the

individual-level predictors: for example, the amount and type of research collaboration undertaken by high-performing researchers compared to other researchers. Past studies also identify three main groups of high citation predictors: author, journal and paper (Walters 2006; Onodera & Yoshikane 2015) and there is some evidence that within these predictors, the most important are those related to the paper (Stewart 1983). Publication analysis can therefore determine whether the papers of high-performing researchers match characteristics of papers known to be associated with high citation rates, such as title length (Elgendi 2019), paper length (Onodera & Yoshikane 2015), number of references (Didegah & Thelwall 2013) and figures (Haslam et al. 2008), coverage of certain topics (Mann et al. 2006) and keywords (Hu et al. 2020), number of authors (Peng & Zhu 2012), quality of journal (Davis et al. 2008), etc. Additionally, predictors relevant to the global North-South divide have been identified in a number of paper-focused studies which found that the author's nationality (whether from the United States or not) (Walters 2006), the regional focus of the articles (focusing on the United States or Europe) and the language of the journal (Van Dalen & Henkens 2005) had a significant positive effect on the average citations.

Three particular theories are employed by these studies to examine the factors affecting publication citation. The *normative view* (Hagstorm 1965; Kaplan 1965; Merton 1973) assumes that science is a normative institution governed by internal rewards and sanctions (Baldi, 1998). According to this perspective the intrinsic characteristics of papers (i.e. content/quality) are the main driver of citations. The *social constructivists' view* (Gilbert 1977; Knorr-Cetina, 1981; Latour 1987) argues that scientific knowledge is socially constructed through the manipulation of political and financial resources and citations are used as persuasion tools (Peng & Zhu 2012). The citing behaviour in this view is driven by a paper's extrinsic characteristics, such as the location of the cited paper author within the stratification structure of science, that would convince the reader with the validity of the arguments (Baldi, 1998). The *natural growth mechanism* (Glänzel & Schoepflin 1995) states that citations are driven mainly by the interaction between publication-level time dependent factors and factors related to the publication outlets. It sees that the characteristics of academic journals (e.g. journal prestige, self-citation rate, maturation speed) will interact with time to affect citations of papers (Peng & Zhu 2012).

All of this supports incorporation of publication-level predictors of outperformance when considering what predicts outperformance of individual researchers. In this study we therefore use a combination of researcher-level data gathering (via interview and survey) and publication-level analysis, in order to provide a fuller picture of research performance predictors.

3.3 Methodology and Data

The positive deviance methodology consists of five steps: "1) *Define* the problem, current perceived causes, challenges and constraints, common practices, and desired outcomes. 2) *Determine* the presence of positive deviant individuals or groups in the community. 3) *Discover* uncommon but successful practices and strategies through inquiry and observation. 4) *Design* activities to allow community members to practice the discovered behaviours. 5) *Monitor* and evaluate the resulting project or initiative" (Positive Deviance Initiative 2010). Time and resourcing constraints meant that only the first three steps could be essayed in this study. We also diverged from the traditional approach by using this study as a testbed for what we will call the "data-powered positive deviance" (DPPD) methodology. Where traditional PD relies on freshly- and specifically-gathered field data, the idea behind DPPD is that it uses pre-existing digital data sources instead of – or in conjunction with – traditional data sources. It uses digital datasets to identify positive deviants (those performing unexpectedly well in a specific outcome measure that is digitally recorded, mediated or observed) and potentially also to understand the characteristics and practices of those PDs if digitally recorded (Albanna & Heeks 2019). The potential of DPPD is that it can mitigate some of the challenges of traditional PD approaches by reducing time, cost and effort, and can improve the positive deviance approach by identifying positive deviants in new or better ways (Albanna & Heeks 2019).

Scientometrics – the field of study that focuses on measuring and analysing scientific literature ('Scientometrics' 2020) – is well suited for a positive deviance approach because, as reflected in the discussion of literature above, research performance does not follow a normal distribution (O'Boyle & Aguinis 2012). Instead, it follows a Pareto or power distribution characterised by strong skewness with a long tail to the right that includes a number of high-performing outliers; sufficient to provide a sample of positive deviants. Scientometrics is well

suited to DPPD specifically for two reasons. First, because the proliferation of electronic research databases has made it possible to develop scientific evaluation indicators that can be used to digitally measure the performance of researchers (e.g. h-index) and journals (e.g. impact factor). Second, because the emergence of advanced data analytics tools alongside the emergence of a variety of large scale datasets (such as citations, references, publication outlets, usage data, paper content, etc.) has made it possible to not only measure performance, but to also analyse the practices of the researchers, characterise their scientific outputs, and predict their future performance. Specifically, this can be rendered possible through techniques such as network analysis, topic modelling, predictive analytics and co-citation analysis.

As discussed above, in this study we used a mixed-methods approach to identify PDs and to analyse their practices. In the *Define* step, we used secondary data from Egyptian university websites and from Google Scholar to set the frame of Egyptian researchers for analysis. In the *Determine* step, we extracted for each researcher the bibliometric data of all his/her academic outputs that were produced while being affiliated to an Egyptian university, without the exclusion of publication or research type. This data was analysed with statistical software R v3.4.1 to identify the positive deviants and non-positive deviants (NPDs) within the overall population. During Stage 1 of the *Discover* step, primary data was collected through in-depth interviews from a sample of PDs to explore practices, attitudes and attributes that might distinguish them from NPDs. During Stage 2 of the *Discover* step, the key findings from Stage 1 plus other predictors of research performance drawn from the literature (see Table 5) were used to design a survey tool. That survey then targeted the whole population and tested if the proposed differentiators were significantly different between the two groups (PDs and NPDs). Finally, in Stage 3 of the *Discover* step, the Scopus secondary dataset was used as the basis for analysis of researcher publications; extending and validating some of the findings identified in the previous steps. Figure 6 summarises the process used to identify PDs and to discover predictors of their performance, and outlines the structure of findings, presented next.

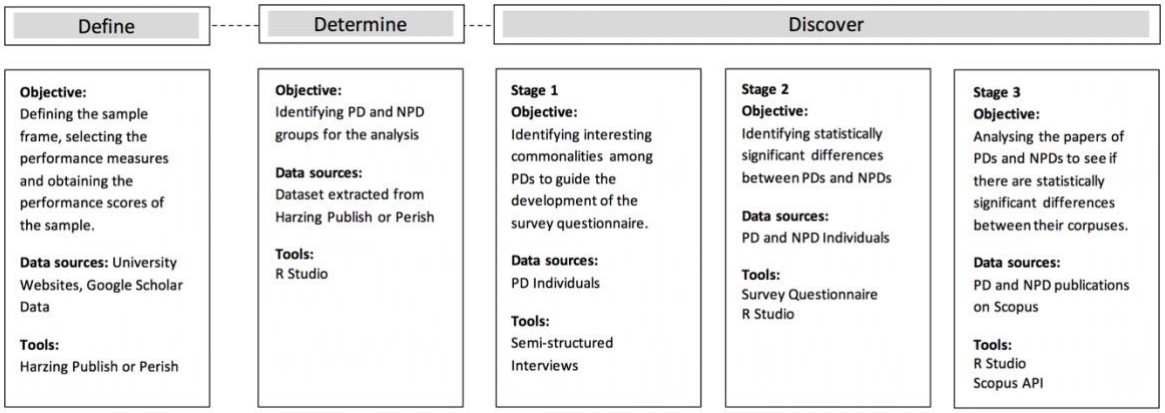

Figure 6: Summary of the applied data-powered positive deviance process

3.4 Findings

3.4.1 Define

The study population comprises IS researchers in Egyptian public universities. A single discipline was chosen to avoid variations in, for example, typical publication and citation rates that arise between different disciplines. Information systems was chosen because of the growing importance of research on digital technologies including technological development and implementation research to economic development, and because a pre-check showed ready presence of a substantial number of Egyptian IS researchers and publications in the main secondary datasets. Egypt was chosen because it was author ***'s home country, with social contacts affording ready access to university departments and staff. Public universities were chosen to ensure that all researchers in the sample worked in a context of similar resource constraints, albeit with slight variations between universities within or outside the Greater Cairo area.

In Egypt there are 29 public universities, 11 of which do not have computer science faculties or IS departments and seven of which do not have an online directory of the IS department staff. So, the final sample included 11 universities: Cairo, Ain Shams, Benha, Helwan, Mansoura, Fayoum, Menofeya, Assiut, Zagazig, Kafrelsheikh and Port Said. The total number of faculty members in those universities was 304 but for this study we only included those

researchers who hold at least a Masters' degree[18] and have published at least one article. This guarantees that they have some publishing experience. Consequently, the final sample that we targeted for this study included 203 researchers who were assistant lecturers, lecturers, assistant professors and professors. (In the Egyptian higher education system, the first academic rank is assistant lecturer, which you receive once you obtain your Masters' degree and then you become a lecturer when you obtain your PhD. The following rank is associate professor and then professor, which are obtained based on both years of experience and publications.)

Using the university websites, the names, degrees and email addresses (if provided) of these researchers were identified and this information was used to extract their citation data from the Publish or Perish (PoP) software. PoP is a freely accessible software program that extracts data for researchers from a number of sources (e.g. Web of Science, Scopus, Google Scholar) to provide a variety of research citation metrics (Harzing 2007). The citation data was extracted for the study sample in July 2018. We only included the papers that had the Egyptian university affiliation i.e. publications produced while doing a PhD abroad were excluded to ensure a fair comparison and to reduce the effects of confounding variables associated with universities abroad.

For this study, Google Scholar was chosen as the source for bibliometrics. The choice of Google Scholar was driven by the fact that ISI citation databases, such as Web of Science, limit citations to journals in the ISI databases. They do not count citations from books and conference proceedings, cover mainly English language articles and provide different coverage in different fields. Such databases significantly underestimate researchers' publications and citations (Ortega 2015). Prior literature also supports the fact that Google Scholar outperforms Web of Science in coverage (Kousha and Thelwall 2007), especially for articles that were published from 1990 onwards (Belew 2005) and for computer science-related research in which conference papers form a key means of publication (Franceschet

[18] It is a prerequisite for confirmed appointment to publish at least two papers towards your Masters' degree in the majority of computer science faculties across the country (those faculties being the typical home of information systems departments and/or researchers).

2010). Additionally, Google Scholar is freely accessible, which makes the DPPD approach used in this case study easily replicable in different scientific fields and countries. The main drawback of Google Scholar is that its consistency and the accuracy of data is lower than commercial citation enhanced databases such as Web of Science (Jacso 2005). Hence, extra time was needed to check the accuracy of obtained results.

For every researcher six main citation metrics – extracted and derived from PoP query results – were used to measure research performance as shown in Table 6.

Citation Metric	Description
h-index	Hirsch's h-index (Hirsch 2005) is the most widely used single-number measure for assessing the research performance of a researcher. It provides a metric that balances impact and productivity. For example, a researcher has a h-index equal to 10 if 10 of his/her papers have received at least 10 citations and the remaining papers received no more than 10 citations.
g-index	Egghe's g-index (Egghe 2006) aims at improving the h-index by giving more weight to highly cited papers. It is calculated based on the distribution of citations received by a given researcher's publications. A g-index of 20 means that an academic has published at least 20 articles that, combined, have received at least 400 citations (i.e. g^2). Unlike the h-index, which requires that each one of the 20 publications should have at least 20 citations, the g-index takes the cumulative number of citations, allowing high numbers to be driven by a small number of articles.
hc-index	The contemporary h-index (Sidiropoulos et al. 2007) rewards academics who maintain a steady level of research activity by giving more weight to recently published articles. The weighting in both the original and the PoP implementations mean, for example, that citations for an article published in the current year count four times whereas citations for papers published four years previously count only once.
hi-index	The individual h-index (Batista et al. 2006) reduces the effects of co-authorship by dividing the standard h-index by the average number of authors in the articles that contribute to the h-index.
aw-index	The aw-index is derived from the age weighted citation rate (AWCR) which measures the number of citations for the articles contributing to the h-index, adjusted for the age of each article, where the count of citations for a specific article is divided by how old it is and then

	summed (Sidiropoulos et al. 2007). The aw-index is defined as the square root of the AWCR to make it more comparable with the h-index. In PoP, the adjusted citation counts are summed across all papers, not just those contributing to the h-index, as this captures the impact of the total body of work more accurately. It also allows more recent and less-cited papers to contribute to the AWCR, even though they might not yet contribute to the h-index.
m-quotient	The m-quotient was proposed by Hirsch to avoid putting early career researchers at a disadvantage (Hirsch 2005) and enabled the inclusion of young researchers in the study sample. It is calculated by dividing the h-index by the publication span (i.e. the number of years since the first publication).

Table 6: Citation metrics extracted for each researcher to measure performance

In this study positive deviants are researchers who outperformed their peers in at least one of the six citation metrics presented in Table 6. The need to use multiple measures was motivated by the drawbacks of relying only on the h-index. These drawbacks include the influence of length of researcher's scientific career, with the h-index reflecting longevity as much as it reflects quality (Alonso et al. 2009; Van Noorden 2010), in addition to its insensitivity to highly cited papers (Egghe 2006). Using multiple measures enabled us to avoid putting certain groups at a disadvantage due to factors such the length of their research career, the size of their research departments, or their publication strategies. Measures like the m-quotient, the aw-index and the hc-index ensured that outperformance is detected regardless of the publication age of the author or the age of the paper. Similarly, some IS departments are larger than others, enabling them to have more research collaborations, and to reduce the potential bias due to the larger pool of research collaborators, the hi-index was employed. We were also interested in researchers with selective publication strategies: those who do not necessarily publish a very high number of papers but who do attain a high impact. This group of researchers can be unfairly assessed using the h-index, while led us to use the g-index. In summary, we can see that these established citation metrics are complementary, as they make different assumptions and have different biases, and that combining the different measures provides a more comprehensive picture of performance.

3.4.2 Determine

Positive deviants are typically identified as specifically-calculated outliers from some measure of central tendency. As can be seen from Figure 7 – violin plots[19] of the distribution of each of the six measures across the entire sample – the data here is not normally distributed. Instead, and consistent with the past findings on researcher performance reported above (O'Boyle and Aguinis 2012) it shows a skewed, Pareto distribution with a long tail above the mean. This makes the *mean* a skewed indicator of central tendency and invalidates the method of identifying positive deviants or outliers in a normally-distributed population, which would define them as those observations lying beyond two or three standard deviations above the mean.

Instead, we used the *median* as an indicator of central tendency and employed the interquartile range (IQR) method (Hampel 1974) to identify positive deviants based on their deviation from the median. In the IQR method, the dataset is divided into four parts, the values that separate the parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively. Q2 is the median of ordered observations, Q1 is the median of observations ordered before Q2 and Q3 is the median of observations ordered after Q2. IQR is Q3-Q1 and outliers are defined as observations that lie beyond 1.5*IQR (Walfish 2006). In this case study, PDs were defined as individuals lying beyond the 1.5*IQR added to the third quartile in at least one of the six citation metrics that we used as measures of performance. In total, 26 unique PDs were identified and their average performance metrics in comparison to the NPDs are summarised in Table 7.

[19] Violin plots are similar to box plots except that they also show the probability density of the data at different values, usually smoothed by a kernel density estimator.

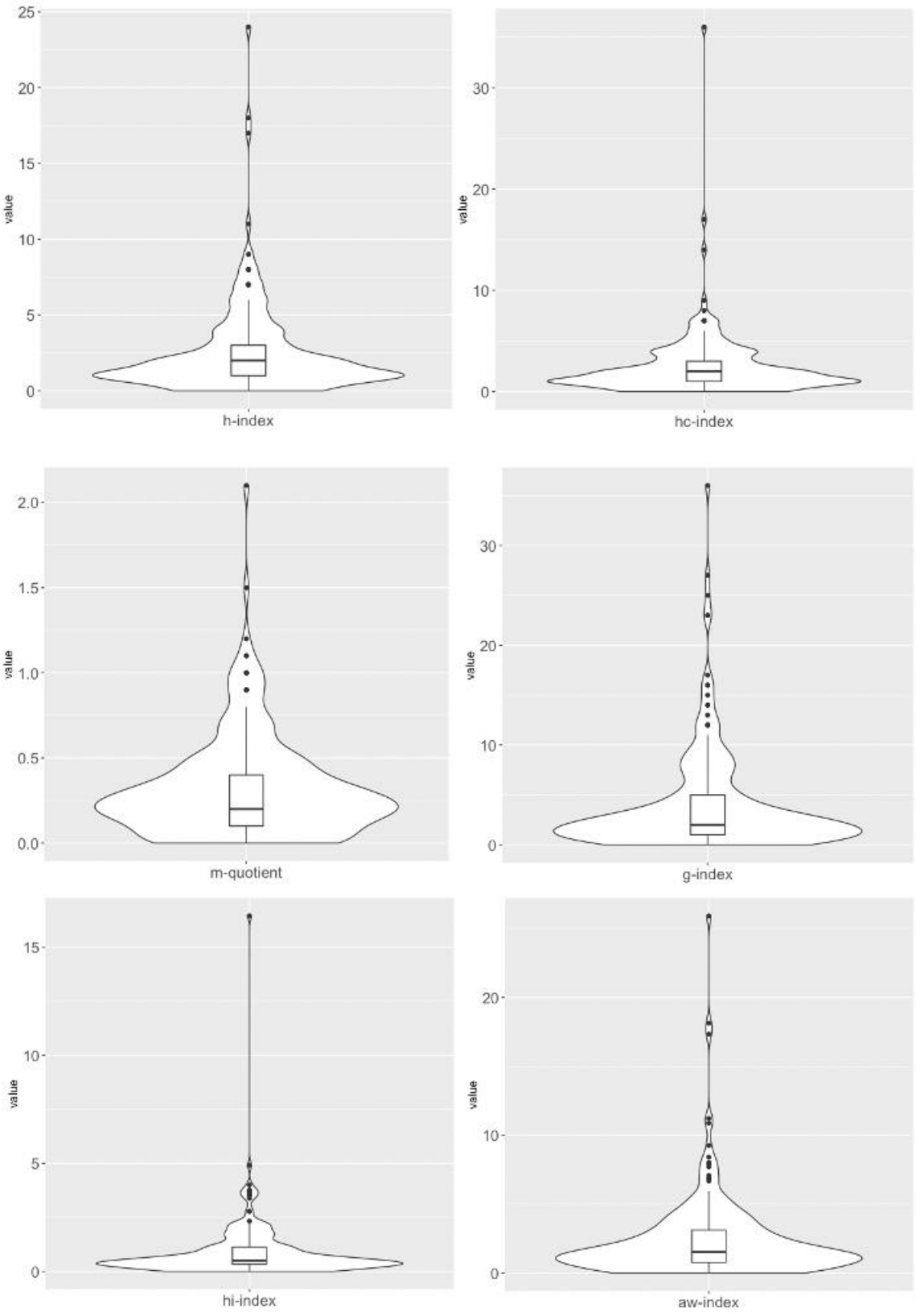
value
h-index
hc-index
m-quotient
g-index
hi-index
aw-index

Figure 7: Violin plots of the sample's scores across the six measures showing the outliers

	PDs (n = 26)	NPDs (n = 177)	Population (n=203)
Average h-index	7.7	1.7	2.5
Average hc-index	7.3	1.7	2.4
Average m-quotient	0.82	0.24	0.3
Average g-index	13.5	2.6	4.0
Average hi-index	2.8	0.58	0.8
Average aw-index	7.8	1.6	2.4
Percentage of assistant lecturers	7.6%	49.2%	43.8%
Percentage of lecturers	26.9%	21.3%	22.2%
Percentage of associate professors	19.2%	18.8%	18.7%
Percentage of professors	46.3%	10.7%	15.3%
Average number of publication years	10.7	6.9	8.3
Average number of papers	43.7	5.9	10.7
Average number of citations	387.1	19.7	66.8

Table 7: Summary statistics of the study population[20]

Cluster Analysis

Hierarchical clustering was used to identify groups of PDs based on the citation metrics in which they were found similar, i.e. all members of a cluster are outliers in similar citation metrics. To support this analysis, a binary vector composed of six dummy variables (representing the six citation metrics) was constructed for each of the positive deviants identified. A value of "1" indicates that this PD is an outlier in the corresponding metric and a value of "0" indicates that this PD is not an outlier in this metric. We then used the *hclust* function of the R *cluster* package[21] to implement complete linkage agglomerative hierarchical clustering using the Gower distance. This method usually yields clusters that are compact and well separated, and the complete linkage criterion ensures direct control of the maximum

[20] As a reminder, these figures exclude papers published while researchers were overseas.

[21] https://cran.r-project.org/web/packages/cluster/cluster.pdf

dissimilarity in each cluster. A graphical representation of the resulting hierarchical tree (i.e. dendrogram) is presented in Figure 8[22].

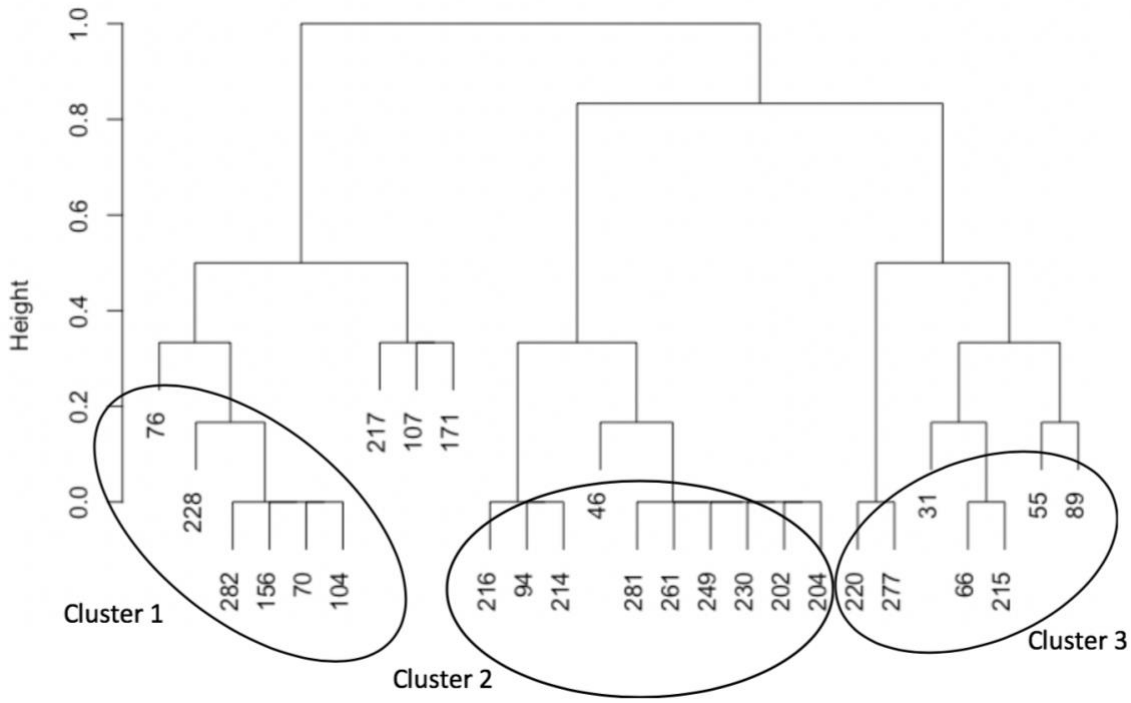

Figure 8: Hierarchical clustering of PD researchers based on their outlier scores

As shown in Figure 8, we were able to cluster the 26 researchers into three main clusters as follows[23].

[22] The numbers in Figure 3 are the unique ID numbers allocated to each individual researcher; allocated from the broad initial population of 304 researchers.

[23] We could not directly group the remaining three researchers (107, 171 and 217) into any of the previous clusters, since each was an outlier in a unique metric, which was not the characterising metric(s) of the identified cluster. Researcher 107 was an outlier in the hc-index, which gives more weight to citations from recent papers. Therefore, this researcher might be considered as one of the "rising stars" despite not being identified as a potential candidate in the dendrogram and having a high publication age (11) in comparison to the average publication age of the group, which is 3.5 as shown in Table 9. Researcher 217 was an outlier in the h-index and researcher 171 was an outlier in the hi-index. Being an outlier in the hi-index indicates independence: researcher 171 has a lot of single-authored papers or papers with a small number of researchers, and his/her average number of authors per paper was only two authors.

Cluster 1: Rising stars

This cluster includes six researchers who were outliers either in the m-quotient, which discounts longevity and citation skews against junior researchers, and/or the aw-index, which gives weight to more recent and as yet less cited papers by calculating age weighted citation rates for the researcher's papers. Researchers belonging to this group were mainly assistant lecturers and lecturers (with the exception of one associate professor) and they were characterised by a short publication span and a small publication volume (as shown in Table 9).

Cluster 2: Exceptional Performers

This cluster includes ten researchers who were outliers in all the six metrics collectively, each being an outlier in at least five metrics. Researcher 46 was an outlier in all metrics except for the hi-index, which might indicate that they have very few single authored papers or that they usually publish with a large number of authors. Researchers 216, 94 and 214 were outliers in all measures except for the m-quotient. The remaining six researchers were outliers in all six metrics. Researchers belonging to this group are characterised by balancing all performance measures i.e. productivity, impact and consistency, and having an old average publication age. They also have the highest average aw-index indicating sustained production of highly cited articles. They were mainly professors with the exception of one lecturer and one associate professor.

Cluster 3: Highly Cited Researchers

This cluster includes seven researchers who are all outliers in the g-index, which gives more weight to highly cited papers. In addition to the g-index, researchers 220 & 227 were outliers in the hi-index which means that they publish mainly individually or with small groups of co-authors; researchers 31 and 89 were outliers in the aw-index, meaning that their highly cited papers are recent, resulting in a high age weighted citation rate. As shown in Table 9, they are characterised by having the longest publication span and the highest number of citations per paper across all clusters.

Table 8 shows the average scores of the three clusters across the six performance measures and Table 9 shows the average scores of those clusters across other relevant performance indicators.

Cluster	m-quotient	h-index	g-index	hc-index	hi-index	aw-index
1	0.95	3	4.5	3.3	0.84	5.13
2	0.98	11.7	20.6	11.9	4.7	11.6
3	0.51	6.8	13.2	5	2.3	6.09

Table 8: Average group scores in the each of the six citation metrics with grouping measures highlighted by colour shading

Cluster	Average publication age	Average no. of papers	Average no. of citations	Average no. of authors/paper	Average no. of cites/paper
1	3.5	6.3	56.1	3.4	8
2	13.3	87.2	726.6	3.0	9
3	14	25	313.85	2.9	16

Table 9: Group scores in relevant performance indicators

3.4.3 Discover

This study used three separate methodologies – in-depth interviews, surveys and publication analysis – to triangulate data on underlying attributes, practices and attitudes of PDs, thus helping to validate findings. The three methodologies are interrelated and were undertaken sequentially in three stages, with the findings from one stage guiding design of the following stage.

Stage 1: Interviews

The objective of this stage was to identify uncommon strategies and practices among positive deviant researchers, which could be used to guide the design of the Stage 2 survey questionnaire. This stage was incorporated into the methodology because predictors of high research performance used in prior studies did not take into account the particular challenges

of global South researchers. Hence there was a need to check the relevance of predictors from past literature and also to explore any additional context-specific predictors.

In order to do this, a semi-structured interview guide (English language version in Appendix A) was developed based on a combination of past literature on high-performing researchers and on the context of global South research. To reduce the need for extensive travel, interviews were restricted to the four universities in Greater Cairo: Cairo, Ain Shams, Helwan and Benha. Those universities were home to 12 of the 26 PDs identified in the *Determine* phase, all of whom were interviewed along with the heads of the IS departments in each university[24]. Table 10 shows the distribution of the interviewed PDs across gender and rank.

Gender	
Male	10
Female	2
Rank	
Assistant Lecturer	1
Lecturer	3
Associate Professor	3
Professor	5
Total Number of Interviews	12

Table 10: PD interviewees across gender and rank measures

Permission was obtained for the interviews to be recorded so that the transcript could subsequently be analysed to identify common themes, patterns and explanations. In all, interview data was coded into nine main categories of potential differentiators of PDs:

Previous education*:* A number of PDs mentioned that they obtained their PhD degrees from global North universities, explaining how it had a fundamental role in changing how they viewed and practised scientific research.

[24] An initial observation was that a number of PDs (n=5 of 26) were also department heads; a predictor that was added to the survey questionnaire.

Research motives: PDs were seen as having different motives and drivers for conducting and publishing research. Getting a promotion was definitely one of those drivers, especially for early-career researchers. However, most of the PDs mentioned motives related to international recognition, staying competitive and how they enjoy the process of publishing research. A number also mentioned that their research satisfies a personal interest they have and publishing in it adds to their satisfaction.

Research type: Almost all of the interviewed PDs worked on applied and experimental research, while only two focused mainly on theoretical research. In terms of topics, all that stood out were areas avoided by most PDs: only one did research in the management of information systems, and only two showed interest in research that had broader social and developmental impact.

Research strategies: A number of PDs said that they were more inclined to do incremental research i.e. building upon previous work; one of them saying *"I do not innovate by finding new problems; I innovate by finding new ways and methods to solve a well-established problem"*. Applying for research funding from schemes like the European Region Action Scheme for the Mobility of University Students (ERASMUS+) and the German Academic Exchange Service (DAAD) was also mentioned by a number of them. Another strategy that was mentioned by most of the PDs is reaching out to foreign authors to conduct collaborative research with them. When they were asked why they do that, their answers varied. Some said that it ensures better access to resources; with sample statements including *"When a paper is accepted in a conference or a journal, their universities can fund their travel expenses or pay for journal submission fees"*, and *"my research requires computing facilities that are hard to provide here and my research partner in Canada has access to those facilities"*. Such resources could include complementary skill sets: *"he [*foreign collaborator*] is good at scientific writing and in the statistical analysis of results and I'm good at coming up with ideas and in the interpretation of results, we were a great team"*. Another view was that foreign authorship assisted publication: *"foreign authors increase the chances of paper acceptance and reduce the time of acceptance drastically"*.

Publication strategies: PDs were aware of the importance of publishing in indexed journals and conferences (such as those indexed in Scopus or ISI), stating it as a major criterion in

selecting where to publish. Within this overall focus, the interviewees' strategies could be grouped into three main categories: a) Publishing in international indexed journals and in international indexed conferences. These interviewees saw top-tier international conferences (e.g. the Very Large Databases series) being as prestigious as journals rated Q1 and Q2 in the SCImago Journal Ranking (SJR)[25] in addition to providing very high paper visibility. b) Publishing in international indexed journals and in local indexed conferences; the majority of the interviewed PDs fell into this category due to financial constraints that limited their ability to attend international conferences. They used conferences as a medium to retain ownership, promote, refine and develop their research ideas before submitting extended versions of papers to journals. *"The journal paper should have at least 30% expansion to the work in the conference paper"* said one of the PD researchers. c) Publishing in international indexed journals only: this group could not afford travel to international conferences and could not find any value in publishing in local conferences stating, for example, that *"Journal papers are more respected in Egypt"*. A number of PDs also stressed the importance of the publisher, indicating that they noticed that there are certain publishers which provided very high visibility to their papers which lead to higher citation. One of them said *"I started to focus on publishers instead of journals, because a strong publisher will make a journal powerful but a strong journal without a strong publisher, will die … any paper published by Elsevier will have great visibility, even if it is a new journal … I would rather publish in a Q4 journal published in Elsevier than publish in a Q3 journal published somewhere else"*.

Research direction: A few PDs mentioned that they usually trace their own citations to see what other authors are saying about their work, and to see how their research is evolving, to get ideas for future work. One of the PDs also mentioned that he follows publishers like ACM, IEEE, Springer and Elsevier to keep informed of new conferences, and thereby indicating hot

[25] "The SCImago Journal Rank (SJR) indicator is a measure of the scientific influence of scholarly journals that accounts for both the number of citations received by a journal and the importance or prestige of the journals where the citations come from." ('SCImago Journal Rank' 2020).

topics, *"if ACM decided to do a conference on recommender systems, this implies that recommender systems are picking up or will become a hot topic"*.

Writing their papers: Interviewees were asked about factors that increase the chances of paper acceptance, and almost all of them agreed on the importance of the paper structure and presentation. One of them even stated that *"a well-written average idea is more likely to be accepted than a poorly-written great idea."* Interviewees also mentioned the importance of issues including: showing the contribution clearly and frequently in the paper, mathematical and theoretical validity, recency of references, use of formal and scientific writing, mastery of the English language, and a self-contained abstract showing clearly the contribution, the method used and the study findings or results. They were also asked about the process of writing a paper but nothing seemed unusual in that regard. Finally, they were asked about factors that could increase paper citations. The answers varied but, again, publishing with a reputable foreign author was mentioned as a key factor in attracting citations. One of the PDs said *"When you are publishing with a trusted author in the field, people feel comfortable to cite his work"*. Publishing in top journals and conferences was also mentioned several times although a few PDs stated that some of their most highly cited work was published in local journals and conferences. PDs also mentioned the importance of publicising research work either through sending emails to researchers they thought would benefit from a paper or through making it available on academic networking sites like ResearchGate and Academia. A number of PDs mentioned the role of the title in attracting citations and one stated that *"I always try to borrow the same keywords used in the titles of the highly cited related papers, because when they search for them, mine will appear."* Some of them also mentioned that survey papers in new fields guarantee high citation and the same for publishing in hot topics at the beginning of their hype cycle.

Research challenges: PDs were asked about their research challenges and how they were able to overcome them. A number of PDs mentioned that they encounter difficulties in choosing the right journal for their publications. Only one PD suggested a way to overcome this, which was through use of online journal finder tools. PDs mentioned the language barrier especially with the students they supervise; one PD stated that he uses the paid-for language editing services provided by Elsevier. Another PD said *"I asked one of my students to stop his PhD for three months just to enhance his English writing by taking courses"*. Some

of them mentioned having overseas contacts that proofread their work. The prolonged time from submission to acceptance was repeatedly mentioned by PDs as a major challenge, especially when the topics they want to publish are time sensitive. In such cases, they would resort to conferences for early communication of those ideas. Finally, all of them mentioned that the limited financial support they receive from the university – to attend conferences and to publish in open access journals – is a major challenge. The alternative was to self-finance their travel and publishing activities or to seek support from funding agencies. Some PDs also mentioned that they overcame this challenge by having as co-authors their former supervisors from their foreign PhD-granting universities, which would sometimes cover conference travel expenses and journal submission fees.

Research skills development: A number of PDs mentioned taking scientific and technical writing courses. One of them also mentioned the importance of formal writing saying that *"I paid a lot of attention to learn the formal way of writing; you learn it from observation, trial and error"*. Indeed, a number mentioned observing highly cited papers written by top authors to see how it is written and structured as a means to enhancing their writing skills. A lot of PDs mentioned using tools like Grammarly *"A number of powerful researchers I know recommended this tool"* said a PD, and Latex, *"I could spend a whole day rearranging figures on Word while it takes me a few minutes using Latex"*.

In summary, the interviews led to the identification of potential patterns in attributes, attitudes and practices amongst PDs: some similar to those from earlier studies but a number that had not previously been identified. Some of these – such as use of keywords from titles of highly cited papers – were practices identified by only one interviewee that, while interesting, were not seen to warrant inclusion in the survey questionnaire. But those appearing repeatedly – studying for a PhD abroad, taking scientific and formal writing courses, publishing with foreign reputable authors, etc – were incorporated into the Stage 2 questionnaire.

Stage 2: Survey

The primary objective of this stage was to validate the findings from the earlier parts of the methodology and to identify predictors of PDs that are significantly different from those of

NPDs. An online survey questionnaire (English language version in Appendix B) was developed based on the review of related work (see e.g. Table 1), amended in light of the findings from Stage 1. A message and link to the survey was sent to the whole sample of PDs and NPDs (n=203) in the 11 universities, including PDs who were interviewed in the previous stage. In total, 90 survey responses were collected: 20 from PDs and 70 from NPDs yielding an overall response rate of 44%.

Survey responses (n=90) were entered and analysed with the use of the statistical software R Studio. Table 11 shows the distribution of the sample of PDs and NPDs across gender and rank. 70% of the respondents held PhD degrees (i.e. were lecturer rank or above) and 30% had MSc degrees (i.e. were assistant lecturers) and the responses came evenly from males and females. It also shows pronounced gender imbalance within the PDs and how seniority still plays a role in being a PD, despite incorporation of measures of performance that would control for that (e.g. hi-index).

	PDs	NPDs
Gender		
Male	17	28
Female	3	42
Rank		
Assistant Lecturer	2	28
Lecturer	5	18
Associate Professor	4	13
Professor	9	11
Total number of Responses	20	70

Table 11: Distribution of the survey responses from PDs and NPDs across gender and rank measures

Feature Selection

The survey tool had 38 questions covering researcher attributes such as gender and rank; attitudes such as what motivates them to publish research; and practices such as the type of research collaborations they engage in. After transforming categorical variables into dummy variables, the final sample had 90 observations and 185 variables. The next step was to build a predictive model to identify significant predictors of PDs among those 185 variables. But

before building such a model, it is important to undertake two necessary steps. The first is to reduce complexity through feature selection i.e. selecting the predictor/independent variables that will be used to predict the dependent variable which in our case was a binary variable with the value of 1 for PD researchers and 0 for NPD researchers. And the second is to identify and address potential issues of multicollinearity.

Feature selection was done by running a simple univariate logistic regression (i.e. relation of the dependent variable with each predictor, one at a time) and then including only predictors that met a certain pre-set cut-off for significance to run in the multiple regression. For the simple regression a cut-off of $p<0.1$ was used since its purpose was to identify potential predictor variables rather than test a certain hypothesis (Ranganathan et al. 2017). A stricter cut-off point ($p<0.05$) was then used in the multiple logistic regression to identify significant predictors of PD. Out of all the explored predictors, 23 were identified as potentially significant predictors as shown in Table 12. Predictors derived from the interviews in Stage 1 are denoted by "(i)".

Following the construction of the simple univariate regression models, we proceeded to check multicollinearity. Specifically, the strength of the association between all possible pairs of 23 predictors was determined using the Spearman rank correlation (for numeric variables), Chi square (for categorical variables) and Anova (for pairs involving one categorical and one numerical variable) implementations in the cor function of the caret package[26]. A lot of the predictors identified were significantly correlated with each other, which would be problematic when jointly used in a multiple logistic regression, creating what is referred to as the separation problem (Mansournia et al. 2018). In practice, stepwise regression can be used to overcome this issue but the problem with this approach is that it might eliminate important predictors that are correlated with the response variable and are important for the user. Partial least squares (PLS) regression allows us to retain in the model all the predictors that have a strong explanatory power. For that reason, it was our preferred method for multiple regression. This is further explained in the following section. PLS regression is a

[26] https://cran.r-project.org/web/packages/caret/caret.pdf

technique that reduces the predictors to a smaller set of uncorrelated components or latent variables and performs least squares regression on these components, instead of performing it on the original predictors. PLS regression is particularly useful when there are more predictors than observations and when the predictors are highly collinear (Abdi 2003).

Predictor	Estimates
Gender	
Male	2.012**
Female	
Marital status	
Married	
Single	
Number of children	
Last Degree	
PhD	
MSc	
Number of years to complete PhD degree	
1-2 years	
2-3 years	
4-5 years	
Foreign PhD degree (i)	0.0938.
Faculty rank	
Ass. Lecturer	
Lecturer	
Ass. Professor	
Professor	1.6946**
Department chair (i)	1.1632*
Supervision	
Undergraduate groups	
MSc students	0.11991*
PhD students	0.18566*
Admin & teaching load	
Publication financial support	
Research grants	1.0468.
College scholarship	
Travel funds	1.2417*
Conference attendance	
Department climate	-0.5078.
Language school	

Work outside university	
Hours of work outside	
Research motivation	
I publish research to get a promotion	
I publish research for international recognition	
I publish research to stay competitive	
I publish research because I enjoy it (i)	
Type of research	
Studies suggesting new ways of viewing/implementing information processing systems e.g. theories, new architectures, new frameworks, ontologies, network protocols	
Research involving the creation of new information-processing systems	
Research involving the creation and evaluation of tools, formalisms, techniques/methods to support existing information processing systems	
Research on social and economic issues related to information processing systems (Including studies of the social and economic impact of information systems, ethical issues, changing views of humanity, etc.)	
From where they get their research ideas	
Publications of researchers I follow on academic platforms (e.g. Google Scholar)	-0.9445.
Live or recorded webinars (e.g. IEEE webinars)	
Papers citing my work (i)	
Predictor	Estimates
Conference attendance	
Future work section of papers	
Research strategy	
I prefer to do radical research	-0.9673.
I prefer to do incremental research (i)	
I prefer to map out broad features of important new areas (i)	
I prefer to probe deeply and thoroughly in narrow areas	
I prefer research which looks for immediate solutions to real life problems (e.g. social problem or industry need)	
I prefer purely theoretical research (i)	
I prefer to carry out research work pretty much on my own	
I prefer to carry out research within a research team	
I prefer long-term projects to short-term ones	
I prefer short-term projects to long-term ones	
Research collaborations	
Doing research with academics in other universities in Egypt	0.4220*
Doing research with academics in other departments in my university	0.4482*

Doing research with academics overseas (i)	0.5985**
Where do you publish research	
Journals indexed in Scopus	
Journals indexed in Thomson ISI (Clarivate Analytics)	
International Conferences with Proceedings indexed in Scopus	
International Conferences with Proceedings indexed in Thomson ISI (Clarivate Analytics)	
Local Indexed Conferences	
Non-indexed Journals	
Non-indexed Conferences	
Factors affecting journal selection	
The publisher of the journal (i)	
Number of issues per year	-0.413*
Editorial board	
Journal fees	
Journal impact factor	
SCImago Journal Rank	
Factors affecting acceptance at top conferences and journals	
Presentation/Structure of the paper (i)	
Reputable co-authors (i)	
Strength of the authors' affiliated universities (i)	
Recency of references (i)	
Including references from the targeted journal/conference proceedings	
Technical depth (i)	
Significance of the contribution (i)	-0.9185*
Theoretical foundation (i)	
Previous publications in the targeted journal/conference	
Primary reason for presenting in conferences	
Interaction with peers and getting feedback	
To be known among my research community	
To publicise my research and attract paper citation	
To gain knowledge about new research areas and trends	
To search for academic posts, possible grants and project collaborations	
Research platforms they use	
Academia	
Semantic Scholar	1.5976.
ResearchGate	
Google Scholar Profile	
Arxiv	
DBLP	

ORCID	
Researcher ID	
ACM	
Publication strategies	
When I start in a new area of research, I prefer publishing the first paper by myself and then including other authors in the following papers (i)	
I publish part of my research work in a conference before publishing it in a journal (i)	
I submit my paper in top conferences (knowing it might get rejected) before submission in journals to get useful feedback/review (i)	
I submit papers in workshops of top conferences (i)	
I publish papers extending/based on the graduation projects of my last year (undergraduate) students (i)	
I publish papers with foreign reputable co-authors (i)	0.4183*
I publish papers in highly ranked journals/conferences (i)	
I publish papers with top publishers (e.g. Elsevier) (i)	
I add my papers in academic networking platforms (e.g. ResearchGate) (i)	
I send hard or soft copies of my paper to researchers in the same field once its published (i)	
I publish papers in specialised journals	
I publish papers in multidisciplinary journals	
I publish papers with new ideas, models or frameworks without experimentation	
I publish papers with new ideas, models or frameworks with experimentation and results	
I publish papers with tools or datasets	
I publish papers in open access journals	
Research challenges	
Motivation to carry out research is a challenge	
Finding the right journal/conference for my paper is a challenge (i)	
Lack of financial support needed for attending conferences is a challenge (i)	
Proficiency of written English is a challenge (i)	
Formal/Scientific writing is a challenge (i)	
Time from submission to acceptance in a journal is a challenge	
Insufficient time because of teaching/admin commitments is a challenge	
Overcoming challenges	
I use journal finder online tools (i)	
I pay for proofreading and editing services for my paper (i)	
I seek external funding agencies (e.g. ITIDA, ASRT, TIEC) to cover the costs of travelling to attend conferences (i)	

I use the financial support provided by the university to cover my travel and publication fees	
I apply for research grants (e.g. Erasmus) (i)	1.2135*
I establish research teams overseas (i)	1.9588**
Research tools	
Grammarly (i)	
Reference managers (e.g. Mendeley)	
Latex (e.g. Sharelatex) (i)	
Enhancing publication quality	
Observing highly cited papers to see how they are written and structured (i)	
English writing courses (i)	1.1386.
Scientific writing / Formal writing courses (i)	1.3458*
Technical courses related to the field (i)	
Using a graphic designer to represent results in an attractive manner (i)	
Sending papers to friends/relatives for proof editing (i)	
Checking paper citations	-1.0756.
To check the geographical distribution of the citing papers	
See the impact of the paper after removing self-citation	
Get ideas on future research areas / improvement areas (i)	

Table 12: Estimated coefficients of significant predictors resulting from the simple logistic regression ***P < 0.001; **P < =0.01; * P < 0.05; '.' P < 0.1. Only those predictors with P<0.1 have their estimates presented in the table.

Multiple Regression

PLS regression is a technique that reduces the predictors to a small set of uncorrelated components and performs regression on those components instead of performing it on the predictors (Tobias 1995). The plsRglm package[27] implements the PLS regression for generalised linear models which is an extension of the classical PLS regression introduced by Bastien et al. (2005). We also used the cv.plsRglm function to identify the ideal number of components to retain in a ten-fold cross-validation (k=10), using six components (nt=6) as

[27] https://cran.r-project.org/web/packages/plsRglm/plsRglm.pdf

the maximum number of components to try with each group or fold. After plotting the results of the cross-validation, we decided to retain only two components based on the mis-classed criterion (i.e. components achieving the least number of misclassified observations) and the non-significant predictor criterion (i.e. components that had significant predictors). Cross-validation with a 70-30 split was used in each of ten training datasets with their test data pairs to calculate the model's prediction accuracy and its AUC i.e. area under the receiver operating characteristics (ROC) curve[28]; which is considered a good metric for evaluating the performance of binary classifiers. Across the ten folds, the model resulted in an average accuracy of 0.78 and an average AUC of 0.70. Significant predictors (p values < 0.05) of the two components we retained are presented in Table 13. The table also shows the estimates of the two retained components across the ten folds with an average coefficient of 1.5 for component one and 1.64 for component two.

Significant Variables	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10	Avg.
Component 1 Estimates	2.2	1.08	2.5	1.13	1.51	1.26	1.45	1.39	1.28	1.17	1.50
Male	0.28	0.38	0.25	0.24	0.40	0.31	0.28	0.30	0.32	0.31	0.31
Professor	0.32	0.32	0.31	0.29			0.30	0.30		0.33	0.31
Foreign PhD degree		0.26								0.26	0.26
Department chair	0.28		0.31	0.29							0.29
Number of supervised MSc students	0.37	0.27	0.32	0.33			0.30	0.35			0.32
Number of supervised PhD students	0.34		0.34	0.33			0.31				0.33
Received research publication grant	0.23										0.23
Received travel funds			0.24				0.25				0.25
Rating of department climate		-0.06					-0.07	-0.05		-0.12	-0.08
I prefer to do radical research that suggests new models frameworks methods and	-0.01						-0.02	-0.06			-0.03

[28] ROC curve is plotted with true positive rate (TPR) against the false positive rate (FPR) where TPR is on the y-axis and FPR is on the x-axis.

architecture that were not implemented before											
The get ideas from publications of researchers they follow on academic platforms e.g. Google Scholar			-0.2								-0.22
Doing research with other academics in other universities in Egypt	0.28		0.23								0.26
Doing research with academics overseas	0.34		0.25	0.34	0.36	0.35	0.34	0.34	0.33	0.34	0.33
Doing research with academics in other departments in my university	0.21		0.21			0.22					0.21
Number of issues per year		-0.29			-0.2		-0.21		-0.21		-0.23
They believe that the significance of the contribution increases the chances of acceptance of a paper in a journal			-0.1			-0.27			-0.24		-0.20
I publish papers with foreign reputable authors	0.33		0.26			0.37	0.33		0.34	0.34	0.33
I apply for research grants				0.25							0.25
I establish research teams overseas	0.24	0.23	0.22	0.24	0.25	0.20	0.24		0.24		0.23
I took English writing courses			0.13					0.18		0.178	0.16
I took Scientific or Formal Writing courses				0.22		0.30			0.25		0.26
They have a profile on Semantic Scholar				0.09							0.09
Component 2 Estimates	2.4	1.13	2.6	1.46	1.33	1.35	1.29	1.29	2.0	1.48	1.64
I prefer to do radical research that suggests											

new models, frameworks, methods and architecture that were not implemented before	-0.36	-0.49			-0.42		-0.02				-0.32
Rating of department climate				-0.52							-0.52
Prediction Accuracy	0.8	0.7	0.9	0.7	0.9	0.9	0.7	0.7	0.8	0.7	0.78
AUC	0.8	0.6	0.8	0.5	0.9	0.8	0.6	0.6	0.7	0.7	0.70

Table 13: Component estimates along with the loadings of their significant predictors and their predictive power in a ten-fold cross-validated PLS model

In the analysis shown in Table 13, significant differences between PDs and NPDs emerged, covering attributes such as gender (PDs were mainly males) and rank (a large number of PDs were professors who are also department chairs). However, it is hard to tell if the latter is a cause or an effect. This is because becoming a department chair in the higher education system in Egypt is mainly based on years of experience rather than academic merit. Additionally, department chairs get the biggest share of MSc and PhD student supervisions, which are also significant predictors of PDs. Having a larger number of students implies a larger number of publications and citations, hence better citation metrics. Differences related to practices included the ways PDs developed their skills, such as taking scientific writing courses and English writing courses and travelling abroad for their PhD degrees. It was also strongly evident that PDs publish more papers with foreign authors. This links to a key difference that persistently appeared, with a relatively high loading, which was doing research with academics overseas. Other collaborations such as doing research with academics in other universities in Egypt and in other departments in the same university, were also significantly higher among PDs but they were not as strong as collaborations overseas. Practices that were found to be less common among PDs included getting research ideas from publications of researchers online, and surprisingly, doing radical research that suggests new models, frameworks, methods and architectures that were not implemented before; which is somewhat counterintuitive. Finally, differences relating to attitudes included how the researchers rated the climate of their department: PDs perceived their departments as more hostile and competitive while NPDs viewed departments as more friendly. They also viewed the number of issues as a less important factor when selecting the journals to publish in.

Table 13 also shows that component one included key predictors that are positively correlated with high research performance while component two was able to better capture the direction of variation of two predictors that are negatively correlated with high research performance ("Rating of department climate" and "I prefer to do radical research that suggests new models frameworks, methods and architecture that were not implemented before") and had very small loadings in component one.

We were also interested in developing a model that would exclude non-controllable factors in order to identify transferable practices that could be adopted by other researchers. It excluded gender, rank and being a department chair. Table 14 presents the findings of this model which resulted in an average accuracy of 0.77 and an average AUC of 0.72. The table also shows the estimates of the two retained components across the ten folds with an average coefficient of 1.55 for component one and 1.58 for component two. Reassuringly, this model's predictive power was very close to the average predictive power of the previous model (mean 0.78 and AUC 0.70) despite the exclusion of those significant predictors that had a relatively high loading. This model reinforces the results from the previous model on the importance of international research collaboration since "Doing research with academic overseas", "Establishing research teams overseas" and "Publishing with foreign reputable authors" appeared repeatedly as significant predictors with high loadings across the ten folds. The significance was also evident of supervising more students (MSc and PhD), having a foreign PhD degree, receiving travel funds, and taking scientific or formal writing courses.

Significant Variables	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10	Avg.
Component 1	2.4	1.38	1.69	1.22	1.64	1.41	1.65	1.42	1.31	1.33	1.55
Foreign PhD degree		0.28								0.30	0.29
Number of supervised MSc students	0.42	0.31	0.37	0.35			0.32	0.38			0.36
Number of supervised PhD students	0.38		0.39	0.35			0.34				0.37
Received research publication grant	0.23			0.27							0.25
Received travel funds			0.27				0.27				0.27

Rating of department climate		-0.10					-0.11	-0.09		-0.18	-0.12
I prefer to do radical research that suggests new models frameworks methods and architecture that were not implemented before	-0.02						-0.05	-0.1			-0.06
The get ideas from publications of researchers they follow on academic platforms e.g. Google Scholar			-0.24								-0.24
Doing research with other academics in other universities in Egypt	0.34		0.27								0.31
Doing research with academics overseas	0.39		0.31	0.39	0.44	0.39	0.40	0.42	0.38	0.41	0.39
Doing research with academics in other departments in my university	0.25		0.23			0.30					0.26
Number of issues per year		-0.34			-0.26		-0.24		-0.22		-0.27
They believe that the significance of the contribution increases the chances of acceptance of a paper in a journal			-0.14			-0.31			-0.26		-0.24
I publish papers with foreign reputable authors	0.38		0.32			0.40	0.37		0.38	0.41	0.38
I establish research teams overseas	0.28	0.27	0.26	0.29	0.32	0.23	0.27		0.28		0.28
I took English writing courses			0.15					0.21		0.21	0.19

I took scientific or formal writing courses				0.25		0.33			0.28		0.29
They have a profile on Semantic Scholar				0.13							0.13
Component 2	2.20	1.31	1.92	1.47	1.31	1.59	1.40	1.15	1.88	1.57	1.58
I prefer to do radical research that suggests new models, frameworks, methods and architecture that were not implemented before	-0.30	-0.52						-0.48			-0.43
Rating of department climate				-0.51							-0.51
Prediction Accuracy	0.7	0.8	0.8	0.7	0.9	0.8	0.7	0.6	0.8	0.78	0.77
AUC	0.8	0.7	0.8	0.5	0.8	0.7	0.7	0.6	0.7	0.74	0.72

Table 14: Component estimates along with the loadings of their significant predictors after excluding gender, rank and role

Stage 3: Publication Analysis

While Stages 1 and 2 were focused on the identification of individual-level predictors of PDs, Stage 3 is focused on the identification of publication-level predictors. In other words, in this stage, the unit of analysis is the paper instead of the researcher. The general motivation for publication-level analysis was noted above but, in addition, some of the significant predictors identified in the previous stages required validation that could only be done through publication analysis. For instance, while PDs mentioned publishing with foreign authors and this was established as a significant predictor, it was not possible to validate this practice and quantify its prevalence within PD publications, relative to NPD publications, without analysing the actual papers. The same was true for the number of authors, the choice of publication outlet, the frequency of research collaborations, etc. In summary, the objective of Stage 3 is twofold: the first is to quantify and validate some of the findings of Stages 1 & 2 through the analysis of the researchers' publications. The second is to identify additional predictors of PDs that can be derived directly from their publications.

In this stage we analysed the publication corpus of PDs versus the publication corpus of NPDs. We defined a PD publication as a paper that has at least one PD author from the 26 high-performing researchers identified in the *Determine* Phase. In contrast, an NPD publication is defined as a paper with at least one NPD author but where none of the authors is a PD. By doing so, we were able to create two mutually exclusive corpora to capture distinguishing characteristics of each. The papers were collected from the Scopus database using the Rscopus[29] data package which links R Studio to the Scopus database API interface. For every researcher in the study population (n=203) a Scopus ID was identified manually through the Scopus advanced search tool form on the website. This ID was then used to retrieve all the possible information associated with their publications including co-authors, co-author affiliations, abstracts, keywords and titles. For consistency purposes, we excluded publications not having the Egyptian university affiliation and/or produced while researchers were abroad (e.g. during overseas PhD study) or produced after 2018 (since the citation metrics upon which we selected the PDs were calculated in 2018). In total, 991 publication records were extracted for PDs and 677 publications were extracted for NPDs. Those publications were further reduced to 876 unique publications (in total), after excluding duplicate publications and publications that did not have abstracts on Scopus. The final corpus of papers included 392 PD papers and 484 NPD papers. Skews consistent with the early-discussed Pareto distribution of performance (O'Boyle and Aguinis 2012) were immediately reflected: PDs make up 13% of the study population but contributed to the creation of 48% of the publications. Those 392 papers were cited 3210 times while NPD papers were cited 1810 times.

We proceeded to examine the three types of paper-level predictors of citation rates used in previous studies: 1) "extrinsic" features of the paper that are not directly related to its content (e.g. paper length, number of authors, etc.); 2) "intrinsic" or content-related features such as the topics covered in the paper; and 3) the publication "outlets" of the paper (e.g. conference or journal paper, journal SJR, etc.). The papers' "extrinsic" features were extracted for each paper using the Scopus API functions. Paper "intrinsic" features were extracted from the

[29] https://cran.r-project.org/web/packages/rscopus/rscopus.pdf

paper title, abstract and keywords using a topic modelling technique that is further explained in a subsequent section. The publication "outlet" features were obtained using the sjrdata[30] package which contains data extracted from the SCImago Journal & Country[31] open data portal.

Publication Predictors

There is a substantial body of research on publication-level predictors of citation rates. In this study we selected several of those predictors based on three main conditions. The first is their relevance to measuring or validating findings from the previous two stages. The second is their relevance to the issues previously raised in the literature relating to Southern researchers. Finally, we excluded features that are difficult to ascertain or would require manual validation (e.g. gender of authors), subjective features (e.g. title attractiveness) or features that would require extensive additional computation or additional measure development e.g. internationalisation of journals. Table 15 presents the different publication features that were used as predictors, and references studies that used them as potential predictors. How paper topics (i.e. paper intrinsic features) were identified and converted into a feature space is explained in the following section.

Predictor	Feature Type	Source
Paper length (number of pages)	Paper-Extrinsic	(Stewart 1983) (Peters & van Raan 1994) (Van Dalen & Henkens 2001) (Walters 2006) (Davis et al. 2008) (Haslam et al. 2008) (Lokker et al. 2008) (Peng and Zhu 2012) (Onodera & Yoshikane 2015) (Elgendi 2019)
Number of authors	Paper-Extrinsic	(Peters & van Raan 1994) (Van Dalen & Henkens 2001) (Walters 2006) (Davis et al. 2008) (Haslam et al. 2008) (Lokker et al. 2008) (Fu & Aliferis 2010) (Peng & Zhu 2012) (Didegah and Thelwall 2013) (Onodera & Yoshikane 2015) (Elgendi 2019)

[30] https://github.com/ikashnitsky/sjrdata

[31] https://www.scimagojr.com/

Intranational affiliations (in Egypt)	Paper-Extrinsic	(Lokker et al. 2008) (Fu & Aliferis 2010) (Didegah & Thelwall 2013) (Onodera & Yoshikane 2015)
International affiliations	Paper-Extrinsic	(Didegah & Thelwall 2013) (Onodera & Yoshikane 2015)
US affiliations	Paper-Extrinsic	(Van Dalen & Henkens 2001)
Type of article	Paper-Extrinsic	(Peters & van Raan 1994) (Van Dalen & Henkens 2001) (Walters 2006) (Davis et al. 2008) (Lokker et al. 2008)
Number of figures	Paper-Extrinsic	(Haslam et al. 2008) (Onodera & Yoshikane 2015) (Elgendi 2019)
Number of references	Paper-Extrinsic	(Stewart 1983) (Peters & van Raan 1994) (Walters 2006) (Davis et al. 2008) (Haslam et al. 2008) (Lokker et al. 2008) (He 2009) (Didegah & Thelwall 2013) (Onodera & Yoshikane 2015)
Title length	Paper-Extrinsic	(Haslam et al. 2008) (Elgendi 2019)
Colons in title	Paper-Extrinsic	(Haslam et al. 2008) (Elgendi 2019)
Abstract length	Paper-Extrinsic	(Lokker et al. 2008)
Paper topics (manual content analysis of popularity and focus)	Paper-Intrinsic	(Peters & van Raan 1994) (Van Dalen & Henkens 2001) (Van Dalen & Henkens 2005) (Lokker et al. 2008) (Peng & Zhu 2012) (Hu et al. 2020)
Open access	Publication Outlet	(Davis et al. 2008)
Publisher	Publication Outlet	This was examined based on insights from the interviews
Journal SJR	Publication Outlet	(Van Dalen & Henkens 2001) (Walters 2006) (Davis et al. 2008) (Haslam et al. 2008) (Fu and Aliferis 2010) (Peng & Zhu 2012) (Didegah & Thelwall 2013)
Publication type (conference paper vs journal article)	Publication Outlet	(Fu & Aliferis 2010)

Table 15: Paper and publication outlet features used as predictors

Topic Extraction

The objective of this analysis is i) to identify the various research topics of the study population, and ii) to develop for every paper a vector representing the distribution of the content across the topics identified. The author topics resulting from this analysis were used as the paper "intrinsic" features in the regression analysis (presented below) in order to explore if there are certain topics that are associated with PD performance and vice versa. Figure 9 explains the steps involved in the topic extraction process. Abstracts are considered a compact representation of the whole article, so we used them as a proxy of the paper content to identify the topics of research. We started by extracting publication data for each author from the paper corpus (876 unique abstracts). Standard text mining pre-processing steps were applied on the entire corpus of abstracts such as lowercasing the corpus, removal of standard stop words (e.g. a, an, and, the), stemming of terms to remove pluralisation or other suffixes and to normalise tenses. Additional pre-processing steps involved removing numbers, special characters and white spaces (Kao & Poteet 2007; Mahanty et al. 2019). An unsupervised topic modelling technique called the Latent Dirichlet Allocation (LDA) (Blei et al. 2003) was used to identify the topics within the abstracts corpus. The basic idea of LDA is that articles will be represented as a mixture of topics, and each topic is characterised by a distribution over words. LDA was applied over the entire corpus to identify topics and calculate the probability distribution across topics for each document.

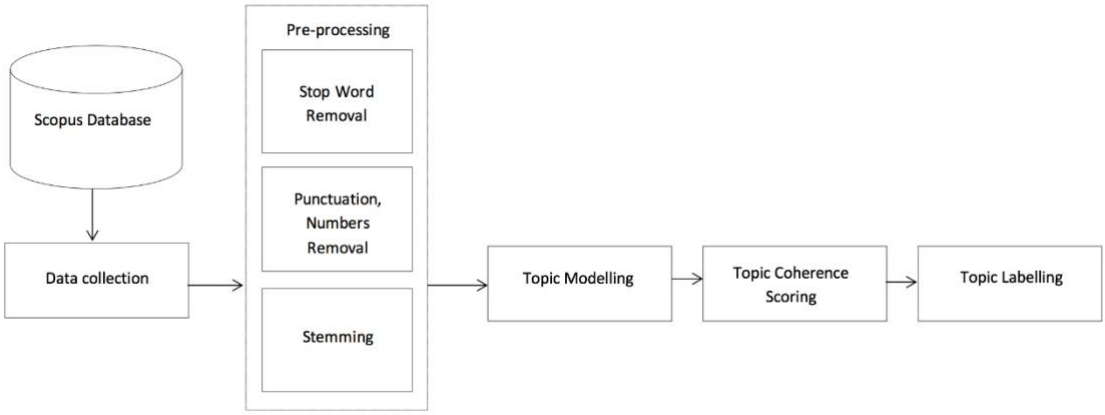

Figure 9: Topic extraction process, developed from Mahanty et al. (2019)

Automatic Topic Coherence Scoring:

To develop the LDA model, we need to have a predetermined value for the number of topics (k). A small number of topics can lead to very generic topics and a large number can result in the generation of overlapping and non-comprehensive topics. Hence, we decided to calculate automatically the topic coherence (the degree of semantic similarity between high scoring words in the topic) at every k from 1 to 20, and established that k=19 achieved the highest topic coherence score as shown in Figure 10.

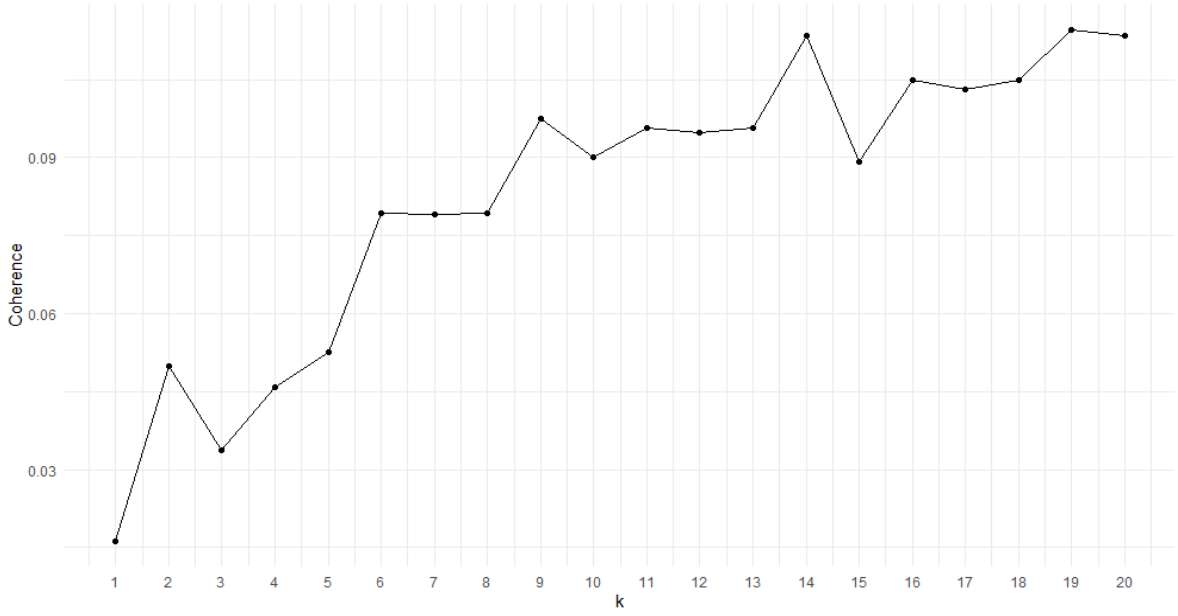

Figure 10: Topic coherence scores

Topic Labelling

Since the labelling of the topics is not done automatically by LDA, we assigned for every topic a relevant label based on the abstracts and keywords of articles with a probability > 90% of falling into that topic. We used the ACM Computing Classification System[32] as the taxonomy

[32] https://dl.acm.org/ccs

that guided the manual assignment of information system labels to topics. We then validated those manually-generated labels by checking if their terms were automatically generated in the most frequent words within that topic. Additionally, expert judgement on the manually assigned labels was achieved by asking the second and third authors to review and approve the labelling performed by the first author. All three co-authors have published research in the information systems field, hence, are able to judge on the validity and relevance of the labels. Table 16 summarises the topics identified, the topic labels assigned and the most frequent keywords.

Topics	Labels	Coherence	Most Frequent Terms
t_1	Neural Networks	0.361	neural, network, neural_network, detect, fast, input, time, result, imag, domain, weight, frequenc, comput, paper, networks, normal, neural_networks, frequenc_domain, number, present
t_2	Distributed Database Management Systems	0.068	data, databas, queri, time, system, propos, process, cloud, distribut, big, result, paper, perform, stream, improv, big_data, storag, update, increas, effici
t_3	Multilevel Programming	0.092	problem, algorithm, optim, propos, solv, object, solut, model, paper, program, level, perform, multi, approach, result, function, genet, genet_algorithm, fuzzi, linear
t_4	Data Mining	0.057	algorithm, propos, graph, cluster, method, similar, protein, paper, object, show, base, data, index, approach, effici, structur, comput, present, mine, mani
t_5	Information Retrieval	0.191	semant, web, arab, search, languag, user, text, retriev, algorithm, inform, propos, document, content, paper, rank, result, extract, generat, approach, model
t_6	Cloud Computing	0.089	secur, data, propos, encrypt, scheme, privaci, key, system, attack, user, imag, access, share, protect, secret, cloud, inform, watermark, digit, util
t_7	Networks	0.222	rout, network, system, protocol, node, propos, base, mobil, blood, biometr, time, traffic, ad_hoc, hoc, keystrok, ad, method, vessel, packet, user
t_8	Web Services	0.103	servic, cloud, mobil, locat, comput, provid, base, communic, propos, web, services, web_servic, user,

			applic, cloud_comput, approach, paper, system, integr, cost
t_9	Wireless Sensor Networks	0.157	network, algorithm, sensor, node, optim, energi, cluster, propos, model, wsn, protocol, wireless, data, power, sensor_network, effici, differ, wireless_sensor, time, springer
t_10	Face Detection	0.069	data, iot, face, neural, process, detect, test, result, comput, time, paper, phase, propos, approach, human, neural_net, net, neural_network, thing, cooper
t_11	Expert Systems & GIS	0.147	decis, gis, make, govern, system, evalu, select, factor, inform, studi, develop, process, problem, criteria, research, success, spatial, decis_make, project, solv
t_12	Product Service Systems & Mobile Based Applications	0.046	learn, technolog, system, agent, busi, framework, paper, complianc, student, differ, present, smart, mobil, process, manag, monitor, bas, support, communic, environ
t_13	Clinical Decision Support Systems	0.086	case, ontolog, base, system, medic, knowledg, bas, propos, cbr, data, health, fuzzi, domain, rough, result, case_bas, model, standard, set, rule
t_14	Hardware Systems	0.084	imag, result, springer, part, optim, nois, signal, state, structur, differ, model, paper, quantum, low, high, method, part_springer, rate, depth, error
t_15	Information System Development Methodologies	0.082	test, softwar, model, system, develop, propos, case, qualiti, approach, requir, paper, test_case, generat, process, design, perform, effort, autom, engin, provid
t_16	Business Process Management & Process Mining	0.003	data, process, approach, model, differ, integr, techniqu, mine, busi, event, exist, paper, work, propos, organ, rule, propos_approach, approaches, analysi, sourc
t_17	Classification Models	0.198	featur, classif, propos, classifi, accuraci, techniqu, differ, appli, result, dataset, select, machin, extract, learn, data, set, approach, algorithm, cancer, support
t_18	Social Network Mining	0.083	user, social, network, recommend, propos, predict, data, base, social_network, detect, opinion, communiti, sentiment, spatial, mine, analysi, system, users, algorithm, show

t_19	Computational Grids	0.037	predict, schedul, system, level, grid, task, result, hcv, signific, wind, studi, patient, risk, time, propos, cell, comput, introduce, respons, resourc

Table 16: LDA generated topics with their corresponding coherence scores and most frequent terms

Time Series Analysis

We were also interested in visualising topic prevalence over time for the PDs and NPDs separately, in a similar way to the analysis presented in the study by Mahanty et al. (2019). This was done by calculating the mean topic proportion per year for the PD corpus and in the NPD corpus as shown in Figure 11. The first finding was that NPDs had a longer publication span starting in 1988 while PDs had a shorter publication span starting in 1993. However, for better visualisation we used the same chronological scale for both groups (2002-2018). There were topics, such as Classification Models, where PDs were early movers and then they were followed by NPDs. We can also clearly see the prevalence of Expert Systems and GIS-related topics in the PD corpus in comparison to the NPD corpus, where there is more prevalence of Neural Networks and Business Process Management & Process Mining. There are also topics that had very similar proportions over time for both groups, such as Social Network Mining.

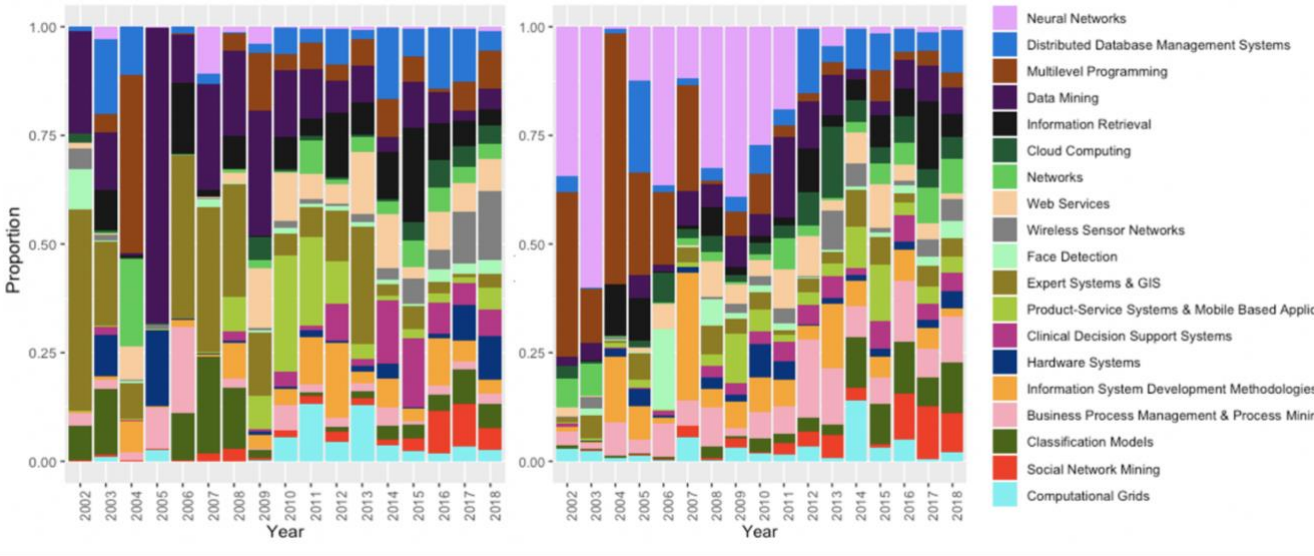

Figure 11: Topic proportions of PD corpus (left) and NPD corpus (right) over time

Feature Selection

Table 17 presents paper features that were used as predictors of PD-authored papers in the multiple logistic regression model. The features with P values less than 0.1 were the ones identified as potential predictors using the same approach adopted in Stage 2. Those 23 features are the ones we selected for the multiple logistic regression model. Subsequently, we calculated all pairwise correlations, using the *cor* function, and we found that a large number of them were correlated. Hence, consistent with Stage 2, we used the PLS regression for generalised linear models as it allowed us to retain all the potential predictors that could have a strong explanatory power. This is further explained in the following section.

Predictor	Estimates
Paper Extrinsic Features	
Paper Length (number of pages)	0.079***
Number of Authors	0.166***
Number of Affiliations (total number of affiliations)	0.119*
Intranational Affiliations (affiliations within Egypt)	
International Affiliations (affiliations overseas)	
US Affiliations (US university affiliations)	0.691*
Type of Paper	
Conference Paper	-0.90***
Journal Article	0.478***
Review Paper	
Number of Figures	
Number of References	0.019.
Title Length	0.072***
Colons in Title	0.467*
Abstract Length	0.012***
Paper Topics or Intrinsic Features	
Neural Networks	-4.77***
Distributed Database Management Systems	
Multilevel Programming	

Data Mining	
Information Retrieval	
Cloud Computing	
Networks	
Web Services	1.25**
Wireless Sensor Networks	2.49***
Face Detection	-1.26*
Expert Systems & GIS	1.31**
Product Service Systems & Mobile Based Applications	
Clinical Decision Support Systems	1.1582*
Hardware Systems	1.64**
Information System Development Methodologies	
Business Process Management & Process Mining	-4.38***
Classification Models	
Social Network Mining	
Computational Grids	
Publication Outlets	
Open Access	
Journal SJR	0.515**
Publisher	
Wiley	1.63**
Springer	0.63***
ACM	-0.68.
Elsevier	0.90***
Inderscience	
IGI Global	
Taylor and Francis	
IEEE	

Table 17: Estimated coefficients of significant predictors resulting from the simple logistic regression ***P < 0.001; **P < =0.01; * P < 0.05; '.' P<0.1. Only those predictors with P<0.1 have their estimates presented in the table.

Multiple Regression

We started by using cv.plsRglm function to identify the ideal number of components to retain in a ten-fold cross-validation (k=10), using six components (nt=6) as the maximum number of components to try with each group or fold. After plotting the results of the cross-validation, we decided to retain only three components based on the mis-classed criterion and the non-significant predictor criterion. Cross-validation with a 70-30 split was used in ten different samples of test and training datasets to calculate the model's prediction accuracy and its AUC. Across the ten folds, the model resulted in an average accuracy of 0.74 and an average AUC of 0.73. Significant predictors (P values < 0.05) within each of the three components were retained and their loadings are presented in Table 18. The table also shows the estimates of the three retained components across the ten folds with an average coefficient of 1.09 for component one and 0.58 for component two and 0.41 for component three.

Significant Variables	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10	Avg.
Component 1	1.14	1.07	1.01	1.07	1.05	1.12	1.15	1.01	1.15	1.14	1.09
Journal Articles	0.34	0.33	0.31	0.35	0.31	0.32	0.31	0.32	0.31	0.33	0.32
Conference Papers	-0.42	-0.41	-0.38	-0.43	-0.39	-0.40	-0.39	-0.40	-0.37	-0.41	-0.40
US Affiliations	0.16	0.16	0.19		0.18	0.22	0.16	0.20		0.18	0.18
Title with Colon					0.08						0.08
ACM_Papers	-0.13				-0.17		-0.17	-0.17	-0.13	-0.12	-0.15
Springer Papers	0.11	0.11	0.07	0.11	0.12	0.07	0.12	0.07	0.11	0.11	0.10
Elsevier Papers		0.18	0.16	0.19	0.18	0.18	0.17		0.11	0.15	0.17
Wiley Papers	0.18	0.16	0.16	0.15	0.18	0.16	0.10	0.21	0.17	0.14	0.16
t_1 Neural Networks	-0.16	-0.18	-0.2	-0.28	-0.16	-0.13	-0.20	-0.14	-0.25	-0.21	-0.19
t_8 Web Services	0.06	0.07	0.09	0.04	0.06		0.08	0.05	0.08	0.06	0.07
t_9 Wireless Sensor Networks	0.23	0.21	0.19	0.24	0.25		0.25	0.20	0.20	0.23	0.22
t_10 Face Detection	-0.03	-0.08					-0.07	-0.07	-0.06		-0.06
t_11 Expert systems & GIS	0.08	0.06	0.07		0.06		0.09	0.08	0.08	0.08	0.08
t_13 Mobile Based Applications		0.09	0.07	0.07	0.08	0.05	0.11	0.06	0.08		0.08

t_14 Hardware systems	0.12	0.15		0.17	0.12	0.14	0.13	0.10	0.13	0.12	0.13
t_16 Business Process Management	-0.17	-0.19	-0.19	-0.18	-0.21	-0.22	-0.25	-0.21	-0.18	-0.17	-0.20
Number of Authors	0.24	0.27	0.26	0.23	0.23	0.26	0.27	0.24	0.24	0.24	0.25
Number of Affiliations	0.27	0.28	0.25	0.25	0.29	0.29	0.28	0.28	0.26	0.27	0.27
Paper Length	0.39	0.40	0.36	0.39	0.38	0.37	0.38	0.39	0.39	0.39	0.38
Number of References	0.26		0.29						0.38		0.31
Title Length	0.21	0.25	0.20	0.20	0.23	0.22	0.21	0.22	0.16	0.21	0.21
Journal SJR	0.29	0.32	0.29	0.30	0.29	0.29				0.29	0.30
Abstract Length	0.35	0.36	0.36	0.38	0.37	0.38	0.40	0.35	0.36	0.35	0.37
Component 2	0.7	0.59	0.5	0.54	0.5	0.7	0.6	0.5	0.52	0.68	0.58
Journal Article	-0.50	-0.42	-0.45	-0.41	-0.44	-0.44	-0.47	-0.37	-0.38	-0.50	-0.44
Conference Paper	0.47	0.41	0.43	0.41	0.43	0.41	0.46	0.34	0.38	0.47	0.42
Elsevier Papers						-0.25					-0.25
t_1 Neural Networks		-0.18									-0.18
t_16 Business Process Management		-0.19		-0.25	-0.27						-0.24
Number of Affiliations	-0.09	-0.13	-0.19	-0.21	-0.14	-0.14	-0.17	-0.10	-0.13	-0.09	-0.14
Paper Length									-0.15		-0.15
Number of References								-0.08			-0.08
Title Length								-0.14			-0.14
Journal SJR	-0.42	-0.36	-0.36	-0.40	-0.39	-0.46	-0.41	-0.36	-0.33	-0.42	-0.39
Abstract Length	0.19	0.20	0.19	0.15	0.17	0.27	0.15	0.17	012	0.35	0.20
Component 3	0.53	0.46	0.41	0.40	0.35	0.35	0.41	0.21	0.43	0.53	0.41
Conference Paper	-0.45	0.43	-0.45	-0.44	-0.47	-0.5	-0.46	-0.36	-0.48	-0.49	-0.37
Prediction Accuracy	0.71	0.76	0.75	0.74	0.7	0.7	0.8	0.71	0.73	0.74	0.74
AUC	0.7	0.76	0.74	0.73	0.7	0.71	0.8	0.7	0.73	0.73	0.73

Table 18: Component estimates along with the loadings of significant predictors and their predictive power in a ten-fold cross-validated PLS model

In the analysis (results shown in Table 18), significant differences emerged between the PD and NPD corpuses. Regarding the paper extrinsic features, it was clear that papers of PDs had longer titles and abstracts and more pages and references. Their papers had a larger number of authors and affiliations which supports the findings of Stage 2 around research collaborations. While Stage 2 showed that PDs are more likely to publish their papers with foreign authors and establish research teams overseas, this analysis enables us to better understand the type of collaborations by showing us that they were mainly with authors from US universities. Additional findings include PDs having more references in their papers and more titles with colons.

Paper intrinsic features, represented by the topics covered in a paper, turned out to be an important distinguishing predictor. It seems that PDs publish fewer papers covering business process management and neural networks in comparison to NPDs. The latter can be linked to an earlier finding from Stage 2 that PDs "do not prefer doing radical research that suggests new models, frameworks, methods and architecture that were not implemented before". One possible explanation could be that neural networks, despite being a cyclical phenomenon, requires radical research whenever there is a recurrence. There were also topics that had much larger coverage in PD papers in comparison to NPD papers, e.g. wireless sensor networks and hardware systems.[33]

As for the publication outlet, we can see that PDs published more journal articles and fewer conference papers; an important predictor that persistently appeared with a high loading. Their preferred publishers were Springer, Elsevier and Wiley, with Elsevier being the one with the highest average loading, supporting the comments of one of the PDs we interviewed who believed that Elsevier journals enable better visibility and impact. PDs were also less likely to publish their papers in ACM. SJR of PD papers was also significantly higher than the SJR of

[33] The significant topics identified in this analysis might not be particularly aligned with the time series analysis conducted earlier (Figure 11) due to the difference in the unit of analysis. The time series compares topic proportions for papers per year, while the regression analysis looks into topic proportions per paper. The former is cumulative, so some topics might look significant cumulatively, such as data mining, but they happen to be insignificant in the regression analysis when topic proportions are analysed individually for each paper.

NPD papers, which implies that PD researchers targeted journals with higher quality and impact.

Table 18 also shows that the first component captures variation associated with PD journal articles while the second component appears to relate to PD conference papers, making it possible to infer some of the characteristics of PD publications in either type of outlet. For instance, component one shows that PD journal articles were correlated with a larger number of authors and affiliations, longer abstracts and higher SJR scores. On the other hand, PD conference papers had fewer affiliations and lower SJR values, while still having long abstracts.

3.5 Discussion and Conclusions

The main motivation of this study was to understand more about research in the global South through a first application of the data-powered positive deviance methodology; a methodology that helped identify and understand those researchers who were able to achieve better research outcomes than their peers. We used a combination of data sources (interviews, surveys and publications) and analytical techniques (PLS regression and topic modelling) to identify predictors of positively-deviant information systems researchers in 11 Egyptian public universities. We found that PDs, despite representing roughly one-eighth (13%) of the study population, contributed to the creation of roughly half (48%) of the publications and achieved nearly double (1.7x) the total number of citations of NPDs.

3.5.1 Significant Predictors of PDs and their Publications

Starting with the practices of PDs, a reasonably clear picture emerged from the analysis showing that significantly more PDs had travelled to get their PhD degrees from global North universities in comparison to NPDs. They had been part of multi-country research teams and published papers with foreign reputable authors. It seems that studying abroad did not just equip them with the technical know-how and the degree needed to pursue their academic careers, but also helped them establish channels of collaboration with their supervisors and their PhD granting universities, long after they returned to their home countries. This confirms findings from previous studies regarding the importance of international research

collaborations (Harris and Kaine 1994; Prpić 1996; Postiglione and Jung 2013; Kwiek 2016; Kwiek 2018). Another significant predictor of PDs was their receipt of research grants and travel funds. The findings also show that PDs took scientific/formal writing and English language courses.

The attitudes of PDs were also different to those of NPDs, when it comes to how they perceived their workplaces. PDs viewed their departments as more hostile or competitive while NPDs viewed them as more friendly. This is somewhat at odds with findings from a previous study that high performers preferred working in a relaxed work environment (Harris and Kaine 1994). Another finding that came as counter-intuitive was that PDs were less inclined to do radical research when compared to NPDs.

In terms of personal attributes, PDs were mainly males and professors, which confirms conclusions from previous studies that identified gender (Prpić 1996; Parker et al. 2010; Patterson-Hazley and Kiewra 2013; Kwiek 2016) and professorship (Kelchtermans and Veugelers 2013; Kwiek 2016) as significant predictors of high performance. A significant number of PDs in comparison to NPDs were department chairs at some point in their academic careers (after becoming professors), which is consistent with a number of studies (Kelchtermans and Veugelers 2013; Patterson-Hazley and Kiewra 2013). PDs also supervised a larger number of postgraduate students, which would help in generation of publications. The ability to select higher quality students would likely result in higher quality publications and more citations, and, in Egypt, department chairs have more leverage than any other academic staff member in the choice of students they will supervise. More generally, the direction of causality here is questionable. For example, given promotions in Egypt are linked to academic performance, it is likely that some of these factors are impacts of above-average research performance; perhaps more so than causes.

While our work did not set out to test particular theories, we can relate findings to all three of the ideas presented earlier. Consistent with the sacred spark notion, PDs clearly did have an internal drive and motivation for undertaking research, though more often mentioned were external rewards or drivers such as promotion, external recognition and competition that fit with utility maximisation. Perhaps most seen was a sense of cumulative advantage

with, for example, researchers who undertook their PhDs overseas then building on that advantage in terms of later publications, grants, and promotions.

The majority of predictors of PD papers, resulting from the publication analysis, are in concordance with existing literature on highly cited papers. They confirm conclusions related to the length of the paper (Onodera and Yoshikane 2015; Elgendi 2019), abstract (Lokker et al. 2008) and title (Haslam et al. 2008); the number of authors (Didegah and Thelwall 2013; Onodera and Yoshikane 2015; Elgendi 2019), co-author affiliations from overseas institutions (Van Dalen and Henkens 2001) and references (Davis et al. 2008; He 2009); and the quality of the journals (Fu and Aliferis 2010; Peng and Zhu 2012). New predictors included the identification of topics that significantly distinguished PDs from NPDs, such as "neural networks" and "wireless sensor networks", along with publishers who were strongly associated with PD papers, such as Elsevier. This thus provides support for theories based around normative, social constructivist and natural growth factors driving publication citation rates.

3.5.2 Predictors Relevant to Global South Challenges

Through this study, we were able to identify predictors of PDs in a global South context that had not been identified in previous studies, and could provide pointers to ways of overcoming challenges specific to Southern researchers. Southern researchers work in contexts of *resource limitation*, and PD researchers apply more for research grants and travel funds from international funding bodies. Some applications included partners from Northern universities, which increased the chances of securing the funds, as those partners are more familiar with grant procurement processes and more experienced in writing proposals. PDs build long-standing research collaborations with their overseas supervisors and PhD granting institutions, which may provide further access to research funds either directly or via joint grant applications. In terms of papers, the publication analysis showed that PDs published more journal articles and fewer conference papers. This choice may relate to seeking profile and citations for outputs: avoiding low-visibility local conferences, and selecting journals as more likely to deliver citations than conferences. But it also fits well as a strategy in the context of limited availability of travel funds. Tendency of PDs to publish with more authors

and with foreign authors could also help pay for journal publication fees, with fees split across more authors or paid from overseas sources.

Southern researchers were seen to encounter *institutional biases* that make it harder for them to get published and cited. PDs are more likely to co-publish with foreign authors, especially US authors, which will help compensate for any such biases among editors, reviewers in single-blind or open review systems, and readers. (Seeking out foreign co-researchers and co-authors also acts as a compensation against the local contextual challenge of there being a *smaller research population* from which to draw research and publication collaborations.) PDs' preference for working on established research areas rather than on radical research topics may also help in relation to institutional barriers, with research that builds incrementally on existing ideas and literature being more likely to be accepted for publication by referees, and cited by others working in the established area. Any biases against citation of work by Southern researchers may be counteracted by PDs' publication of papers with more authors and more affiliations than NPDs. Having multiple authors and affiliations increases the likelihood of citations, as each author has their own network and bringing those networks together can increase readership (Elgendi 2019). Multiple authorship may also enrich the paper through the integration of different perspectives and expertise, which could lead to greater citation (Peng & Zhu 2012). Similarly, PDs publish papers with a larger number of references which increases paper visibility through citation-based search in databases that allow it, such as Google Scholar (Didegah and Thelwall 2013), and through the "tit-for-tat" hypothesis i.e. authors tend to cite those who cite them (Webster et al. 2009).[34] By and large, then, this tends to support social constructivist views of publication citations; showing how contextual factors influence publication – but also how researchers seek to compensate when those factors may tend to reduce citation rates.

Southern researchers work in contexts of *lower English proficiency*, and PDs were shown to take scientific writing and English writing courses more than NPDs, and their greater

[34] More references might also indicate more comprehensive work, hence a better quality paper, and could mean a large related field, hence better citations (Moed et al. 1985).

likelihood of PhD study at a global North university may also have enhanced their command of English.

3.5.3 Methodological Innovation

The use of six different citation metrics enabled us to evaluate performance using different dimensions while controlling for factors that could disadvantage certain groups. It also enabled us to identify and profile PD researchers into three main clusters: rising stars, high performers and highly cited researchers. It was not possible to investigate predictors specific to each cluster individually, due to their small sample size, but this could be a possible avenue for future research.

The majority of studies on predictors of high-performing researchers have focused on individual-level and institution-level predictors. This is one of very few studies that examined publication-level predictors along with individual-level predictors through multiple stages, angles and by triangulating different sources of data. This multi-stage process was both useful and insightful. A number of assumptions about PDs in Stage 1 turned out to be not statistically significant in Stage 2. Examples include practices related to research publishing like submitting papers in conferences followed by extended submissions in journals, paying for proofreading services, and practices related to where researchers publish their work. On the other hand, Stage 1 was crucial because it led to the discovery of PD predictors (some of which had not previously been examined in the literature) that proved to be significant in the statistical analysis that was conducted in Stages 2 and 3. Examples of those predictors include but are not limited to publishing with foreign reputable authors, taking scientific and formal writing courses, and the selection of journal publishers.

Stage 3 enabled us to better understand significant predictors that were identified in Stage 2, e.g. the discovery that teams established overseas were mainly located in the US with authors having US university affiliations. Stage 3 was also useful in quantifying the types of papers (i.e. conference paper, review paper and journal article), and the quality of journals and their different publishers. Although recent studies already demonstrated that topic-related paper features increase the predictive power of highly cited research (Hu et al. 2020), this is one of very few studies that combined topic or paper intrinsic features with extrinsic and publication

outlet features to predict papers of high performers. We also explored the adoption and prevalence of topics over time in each of the PD and NPD corpora, which provided additional longitudinal insights that were not possible to capture through the regression analysis alone. It also emerged that it was possible to predict a paper of a PD from its features with an accuracy that is similar to predicting if this researcher is a PD using his/her survey response.

Through this study, we demonstrated that application of the DPPD methodology has potential value to the scientometrics field. Advances in this field have enabled digital measurement and tracking of researchers' performance using multiple dimensions, and the open nature of their digital products (i.e. publications) enabled us to digitally quantify and identify some of their publication strategies and research directions. DPPD also provided means to reduce the qualitative search space by limiting the interviewing to a smaller sample of information rich individuals i.e. PDs, thus reducing the time needed for hypothesis generation. Finally, the "data powered" aspect of DPPD characterised by combining digital data with traditional data helped us confirm and better understand the identified predictors.

3.5.4 Practical Implications

The key finding of this study is the identification of a set of factors that are significant predictors of PD outcomes. Our analysis cannot, of course, guarantee that applying these factors more broadly would lead to the same outcomes achieved by PDs. Additionally, although causal connections have been outlined in many instances, correlates of high performance do not necessarily imply a causal relation. The work here has only covered one academic discipline in one global South location: replication in other disciplines and countries represents a future research agenda.

One must also step back and recognise two things. First, that citation-based research performance is not the "be all and end all" that should mechanically shape research: relevance of topic to national socio-economic challenges or development of Southern-based methodology and theory could also, for example, be important criteria for Southern researchers. Second, that the findings here in part reflect structural impediments. The fact that highly-cited researchers are overwhelmingly male would not, for example, generate the

practical implication that there should be greater resource flows to men in order to generate stronger research performance!

Nonetheless, there would be value in individual Southern researchers reflecting on the research- and paper-related behaviours that have been shown associated with positive-deviant research profiles. These include publishing with multiple authors from different institutions (domestic and international); establishing connections with foreign reputable authors; including a large number of references; having a comprehensive abstract; publishing in particular journals instead of conferences; and contributing to mainstream topics that build on existing work.

Higher education institutions and higher education policy makers may also reflect on the findings, and consider strategic implications for training, resource provision, collaborations, etc. For example, English and scientific/formal writing courses were associated with PD performance; such training could be part of the mandatory training that academics are required to take in order to be promoted in the Egypt's higher education system. Training could be designed around research grant writing and providing guidance on funding bodies that researchers can apply to. International research collaborations appeared as an important predictor of PDs; so, university senior managers and policy makers can explore ways to reduce barriers and increase opportunities for overseas PhD study, post-PhD return, and ongoing joint research projects with global North universities.

3.6 Future Research

This study has developed and tested a methodology that could be replicated in other contexts, such as other countries or other academic disciplines. However, it only covered the first three stages of the DPPD method: defining the problem, determining positive deviants, and discovering the PD practices and strategies. The last two stages of the DPPD method concerned with designing and implementing interventions, and monitoring and evaluating their effects on the intervention population, were not included in this study due to time and resource constraints. There is an opportunity for future research to apply the full DPPD method especially given that the performance indicators that were captured for the study population could relatively easily be used for monitoring and evaluating interventions.

Furthermore, this study demonstrated how the different citation metrics enabled us to cluster and profile researchers into certain groups. However, there is still an opportunity to explore cluster-specific predictors of performance if the sample size per cluster/group is big enough to infer potential hypotheses related to members of the group. Those predictors can then inform cluster-specific interventions. For example, identifying predictors specific to the 'rising stars' group (characterised by the low publication age) and using those findings to design interventions targeting young researchers. Additional publication-level predictors, such as the data on author contributions increasingly available via initiatives such as CRediT, can be used to understand how collaboration takes place in papers of positive deviants versus papers of non-positive deviants. Future research could also include network analysis using co-authorship data to investigate the relationship between research groups and PD performance, and to measure the magnitude of local and international collaboration that positive deviants engage in.

In this study we looked into individual-level and publication-level predictors, but we did not look into institutional-level predictors which could provide a more holistic understanding of outperformance. Looking into such supra-individual factors could uncover potential structural factors that are either conducive to or hinder outperformance within institutions. Identifying such factors could then inform higher education policies.

References

Abdi, H. (2003) Partial least square regression (PLS regression), *Encyclopaedia for Research Methods for the Social Sciences*, *6*(4), 792–795.

Albanna, B., & Heeks, R. (2019) Positive deviance, big data, and development: A systematic literature review, *Electronic Journal of Information Systems in Developing Countries*, *85*(1), e12063.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009) h-Index: A review focused in its variants, computation and standardization for different scientific fields, *Journal of Informetrics*, *3*(4), 273–289.

Altanopoulou, P., Dontsidou, M., & Tselios, N. (2012) Evaluation of ninety-three major Greek university departments using Google Scholar, *Quality in Higher Education*, *18*(1), 111–137.

Baldi, S. (1998) Normative versus social constructivist processes in the allocation of citations: A network-analytic model, *American Sociological Review, 63*(6), 829-846.

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005) PLS generalised linear regression, *Computational Statistics & Data Analysis*, *48*(1), 17–46.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006) Is it possible to compare researchers with different scientific interests? *Scientometrics*, *68*(1), 179–189.

Belew, R. K. (2005) Scientific impact quantity and quality: Analysis of two sources of bibliographic data, *arXiv:cs.IR/0504036 v1*.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003) Latent dirichlet allocation, *Journal of Machine Learning Research*, *3*, 993–1022.

Blicharska, M., Smithers, R.J., Kuchler, M., Agrawal, G.K., Gutiérrez, J.M., Hassanali, A., Huq, S., Koller, S.H., Marjit, S., Mshinda, H.M. and Masjuki, H. (2017) Steps to overcome the North-South divide in research relevant to climate change policy and practice, *Nature Climate Change*, *7*(1), 21–27.

Cole, J.R. & Cole, S. (1973) *Social Stratification in Science*. Chicago, IL: University of Chicago Press.

Confraria, H., Godinho, M. M. & Wang, L. (2017) Determinants of citation impact: A comparative analysis of the Global South versus the Global North, *Research Policy*, *46*(1), 265-279.

Copes, H., Khey, D. N. & Tewksbury, R. (2012) Criminology and criminal justice hit parade: Measuring academic productivity in the discipline, *Journal of Criminal Justice Education*, *23*(4), 423–440.

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G. & Connolly, M. J. L. (2008) Open access publishing, article downloads, and citations: randomised controlled trial, *British Medical Journal*, *337*, a568.

Didegah, F. & Thelwall, M. (2013) Determinants of research citation impact in nanoscience and nanotechnology, *Journal of the American Society for Information Science and Technology*, *64*(5), 1055–1064.

Egghe, L. (2006) Theory and practise of the g-index, *Scientometrics*, *69*(1), 131–152.

Elgendi, M. (2019) Characteristics of a highly cited article: a machine learning perspective, *IEEE Access*, *7*, 87977–87986.

Flanigan, A. E., Kiewra, K. A. & Luo, L. (2018) Conversations with four highly productive German educational psychologists: Frank Fischer, Hans Gruber, Heinz Mandl, and Alexander Renkl, *Educational Psychology Review*, *30*(1), 303–330.

Franceschet, M. (2010) A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar, *Scientometrics*, *83*(1), 243–258.

Fu, L. D. & Aliferis, C. F. (2010) Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature, *Scientometrics*, *85*(1), 257–270.

Glänzel, W. & Schoepflin, U. (1995) A bibliometric study on ageing and reception processes of scientific literature, *Journal of Information Science*, *21*(1), 37–53.

Gibbs, W. W. (1995) Lost science in the third world, *Scientific American*, *273*(2), 92–99.

Gilbert, G. N. (1977) Referencing as persuasion, *Social Studies of Science*, *7*(1), 113-122.

Goldemberg, J. (1998) What is the role of science in developing countries? *Science*, *279*, 1140–1141.

Gonzalez-Brambila, C. N., Reyes-Gonzalez, L., Veloso, F. & Perez-Angón, M. A. (2016), The scientific impact of developing nations, *PLoS One*, *11*(3), e0151328.

Hagstrom, Warren o. (1965) *The Scientific Community.* New York: Basic Books.

Hampel, F. R. (1974) The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, *69*, 383–393.

Harris, G. & Kaine, G. (1994) The determinants of research performance: A study of Australian university economists, *Higher Education*, *27*(2), 191–201.

Harzing, A. W. (2007) *Publish or Perish*. http://www.harzing.com/pop.htm

Haslam, N., Ban, L., Loughnan, S., Peters, K., Whelan, J. & Wilson, S. (2008) What makes an article influential? Predicting impact in social and personality psychology, *Scientometrics*, *76*(1), 169–185.

He, Z. (2009) International collaboration does not have greater epistemic authority, *Journal of the American Society for Information Science and Technology*, *60*(10), 2151–2164.

Hirsch, J. E. (2005) An index to quantify an individual's scientific research output, *Scientometrics, 85(3),* 741-754.

Hu, Y., Tai, C., Ernest, K. & Cai, C. (2020) Identification of highly-cited papers using topic-model-based and bibliometric features: the consideration of keyword popularity, *Journal of Informetrics*, *14*(1), 101004.

Jacso, P. (2005) As we may search - comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases, *Current Science*, *89*(9), 1537-1547.

Kao, A. & Poteet, S. R. (2007) *Natural Language Processing and Text Mining*. Washington, DC: Springer Science & Business Media.

Kaplan, N. (1965) The norms of citation behavior: Prolegomena to the footnote, *American Documentation*, *16*(3), 179-184.

Karlsson, S., Srebotnjak, T. & Gonzales, P. (2007) Understanding the North-South knowledge divide and its implications for policy: a quantitative analysis of the generation of scientific knowledge in the environmental sciences, *Environmental Science and Policy*, *10*(7–8), 668–684.

Kelchtermans, S. & Veugelers, R. (2013) Top research productivity and its persistence, *Review of Economics and Statistics*, *95*(1), 273–285.

Khey, D. N., Jennings, W. G., Higgins, G. E., Schoepfer, A. & Langton, L. (2011) Re-ranking the top female academic 'stars' in criminology and criminal justice using an alternative method: A research note, *Journal of Criminal Justice Education*, *22*(1), 118–129.

Kiewra, K. A. & Creswell, J. W. (2000) Conversations with three highly productive educational psychologists: Richard Anderson, Richard Mayer, and Michael Pressley, *Educational Psychology Review*, *12*(1), 135–161.

King, D. A. (2004) The scientific impact of nations, *Nature*, *430*(6997), 311–316.

Knorr-Cetina, K. (1981) *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science.* Oxford, UK: Pergamon Press.

Kousha, K. & Thelwall, M. (2007) Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis, *Journal of the American Society for Information Science and Technology*, *58*(7), 1055–1065.

Kwiek, M. (2016) The European research elite: a cross-national study of highly productive academics in 11 countries, *Higher Education*, *71*(3), 379–397.

Kwiek, M. (2018) High research productivity in vertically undifferentiated higher education systems: Who are the top performers? *Scientometrics*, *115*(1), 415–462.

Kyvik, S. (1990) Age and scientific productivity. Differences between fields of learning, *Higher Education*, *19*(1), 37–55.

Latour, B. (1987) *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, MA: Harvard University Press.

Leimu, R. & Koricheva, J. (2005) What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*, *20*(1), 28–32.

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L. & Haynes, R. B. (2008) Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study, *British Medical Journal, 336*(7645), 655–657.

Mahanty, S., Boons, F., Handl, J. & Batista-Navarro, R. (2019) Studying the evolution of the 'circular economy' concept using topic modelling. In *International Conference on Intelligent Data Engineering and Automated Learning*. Cham: Springer, 259–270

Man, J. P., Weinkauf, J. G., Tsang, M. & Sin, D. D. (2004) Why do some countries publish more than others? An international comparison of research funding, English proficiency and publication output in highly ranked general medical journals, *European Journal of Epidemiology*, *19*(8), 811–817.

Mann, G. S., Mimno, D. & McCallum, A. (2006) Bibliometric impact measures leveraging topic analysis. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries - JCDL '06*. New York, NY: ACM Press, 65–74.

Mansournia, M. A., Geroldinger, A., Greenland, S. & Heinze, G. (2018) Separation in logistic regression: causes, consequences, and control, *American Journal of Epidemiology, 187*(4), 864–870.

Martínez, R. S., Floyd, R. G. & Erichsen, L. W. (2011) Strategies and attributes of highly productive scholars and contributors to the school psychology literature: Recommendations for increasing scholarly productivity, *Journal of School Psychology*, *49*(6), 691–720.

Mayrath, M. C. (2008) Attributions of productive authors in educational psychology journals, *Educational Psychology Review*, *20*(1), 41–56.

Merton, R. K. (1968) The Matthew effect in science: The reward and communication systems of science are considered, *Science*, *159*(3810), 56–63.

Merton, R. K. (1973) *The Sociology of Science: Theoretical and Empirical Investigations*, Chicago, IL: University of Chicago Press.

Moed, H. F., Burger, W. J. M., Frankfort, J. G. & Van Raan, A. F. J. (1985) The use of bibliometric data for the measurement of university research performance, *Research Policy, 14*(3), 131–149.

National Science Board (2018) *Science and Engineering Indicators 2018*. NSB-2018-1. Alexandria, VA: National Science Foundation.

O'Boyle, E. & Aguinis, H. (2012) The best and the rest: Revisiting the norm of normality of individual performance, *Personnel Psychology*, *65*(1), 79–119.

Onodera, N. & Yoshikane, F. (2015) Factors affecting citation rates of research articles, *Journal of the Association for Information Science and Technology*, *66*(4), 739–764.

Ortega, J. L. (2015) How is an academic social site populated? A demographic study of Google Scholar Citations population, *Scientometrics*, *104*(1), 1–18.

Parker, J. N., Allesina, S., & Lortie, C. J. (2013) Characterizing a scientific elite (B): publication and citation patterns of the most highly cited scientists in environmental science and ecology, *Scientometrics*, *94*(2), 469–480.

Parker, J. N., Lortie, C. & Allesina, S. (2010) Characterizing a scientific elite: the social characteristics of the most highly cited scientists in environmental science and ecology, *Scientometrics*, *85*(1), 129–143.

Pasgaard, M. & Strange, N. (2013) A quantitative analysis of the causes of the global climate change research distribution, *Global Environmental Change*, *23*(6), 1684–1693.

Patterson-Hazley, M. & Kiewra, K. A. (2013) Conversations with four highly productive educational psychologists: Patricia Alexander, Richard Mayer, Dale Schunk, and Barry Zimmerman, *Educational Psychology Review*, *25*(1), 19–45.

Peng, T. Q. & Zhu, J. J. H. (2012) Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies, *Journal of the American Society for Information Science and Technology*, *63*(9), 1789–1803.

Peters, H. P. F. & van Raan, A. F. J. (1994) On determinants of citation scores: A case study in chemical engineering, *Journal of the American Society for Information Science*, *45*(1), 39–49.

Positive Deviance Initiative (2010) *Basic Field Guide to the Positive Deviance Approach*. Boston, MA: Tufts University.

Postiglione, G. A. & Jung, J. (2013) World-class university and Asia's top tier researchers. In Wang, Q, Ying C, Cai Liu N, eds. *Building World-Class Universities: Different Approaches to a Shared Goal*. Rotterdam: SensePublishers, 161–179.

Prpić, K. (1996) Characteristics and determinants of eminent scientists' productivity, *Scientometrics*, *36*(2), 185–206.

Ranganathan, P., Pramesh, C. & Aggarwal, R. (2017) Common pitfalls in statistical analysis: logistic regression, *Perspectives in Clinical Research*, *8*(3), 148–151.

Salager-Meyer, F. (2008) Scientific publishing in developing countries: Challenges for the future, *Journal of English for Academic Purposes*, *7*(2), 121–132.

'Scientometrics' (2020). *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Scientometrics

'SCImago Journal Rank' (2020). *Wikipedia*. Available at: https://en.wikipedia.org/wiki/SCImago_Journal_Rank

Sidiropoulos, A., Katsaros, D. & Manolopoulos, Y. (2007) Generalized Hirsch h-index for disclosing latent facts in citation networks, *Scientometrics*, *72*(2), 1–34.

Sternin, J. (2002) Positive deviance: a new paradigm for addressing today's problems today, *The Journal of Corporate Citizenship*, *5*, 57–62.

Sternin, M., Sternin, M. D., & Marsh, D. (1997) Rapid, sustained childhood malnutrition alleviation through a positive deviance approach in rural Vietnam: Preliminary findings. In Wollinka O, Keeley E, Burkhalter RB, Bashir N, eds. *The Hearth Nutrition Model: Applications in Haiti, Vietnam, and Bangladesh*. Arlington, VA: BASICS, 49–61.

Stewart, J. A. (1983) Achievement and ascriptive processes in the recognition of scientific articles, *Social Forces, 62*(1), 166–189.

Thesee, G. (2006) A tool of massive erosion: Scientific knowledge in the neo-colonial enterprise. In Sefa Dei GJ & Kempf A, eds. *Anti-Colonialism and Education*. Rotterdam: Sense Publishers, 25–42.

Tobias, R. D. (1995) An introduction to partial least squares regression. *Proceedings of the Twentieth Annual SAS Users Group International Conference* (Vol. 20). Cary, NC: SAS Institute, 1250–1257.

Van Dalen, H. P. & Henkens, K. (2001) What makes a scientific article influential? The case of demographers, *Scientometrics*, *50*(3), 455–482.

Van Dalen, H. P. & Henkens, K. (2005) Signals in science – On the importance of signaling in gaining attention in science, *Scientometrics*, *64*(2), 209–233.

Van Noorden, R. (2010) A profusion of measures: scientific performance indicators are proliferating--leading researchers to ask afresh what they are measuring and why, *Nature*, *465*(7300), 864–866.

Walfish, S. (2006) A review of statistical outlier methods, *Pharmaceutical Technology, 30*(11), 82.

Walters, G. D. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact. *Scientometrics*, *69*(3), 499–510.

Webster, G. D., Jonason, P. K. & Schember, T. O. (2009) Hot topics and popular papers in evolutionary psychology: Analyses of title words and citation counts in evolution and human behavior, 1979 – 2008, *Evolutionary Psychology*, *7*(3), 147470490900700301.

White, C. S., James, K., Burke, L. A. & Allen, R. S. (2012) What makes a 'research star'? Factors influencing the research productivity of business faculty, *International Journal of Productivity and Performance Management*, *61*(6), 584–602.

World Bank (2020) *Science & Technology Indicators*. Washington, DC: World Bank.

Yair, G., Gueta, N. & Davidovitch, N. (2017) The law of limited excellence: publication productivity of Israel Prize laureates in the life and exact sciences, *Scientometrics*, *113*(1), 299–311.

Appendix

Appendix A: Stage 1 Interview Guide

The interview guide was divided into two levels. Level two questions (L2Q) were questions for which answers were sought i.e. via mental inquiry; level one questions (L1Q) were questions addressed to the interviewee directly i.e. via verbal inquiry.

L2Q1: Do PDs have different motives?

L1Q1: What motivates you to publish your research?

L2Q2: Do PDs publish different types of research?

L1Q2: What kind of research do you prefer? (E.g. review, model development, coding, data analysis, etc)

L2Q3: Do PDs have different research strategies?

L1Q3.1: Where do you usually publish your research? (E.g. local conferences, international conferences, local journals, international peer reviewed journals, etc)

L1Q3.2: Are there specific conferences that you always attend?

L1Q3.3: What kind of co-authorship do you prefer the most? (E.g. student, colleague, someone from the department, someone outside the department, international co-authors)

L2Q4: How do PDs increase the chances of paper acceptance?

L1Q4.1: How do you decide on research material that qualifies for publication?

L1Q4.2: How do you decide if this is a conference paper or a journal article?

L1Q4.3: Are there specific conferences you target for paper submission?

L1Q4.4: Are there specific journals you target for paper submission?

L2Q5: Do PDs perform certain practices that increase paper citation as per the theoretical prepositions found in the literature?

L1Q5.1: How do you plan to write a paper? How long does it take to publish a paper? Are there certain steps that you perform for research publication?

L1Q5.2: What constitutes your literature review?

L1Q5.3: Do you target a minimum number of references in your articles?

L1Q5.5: How do you select the papers which you will cite or use as your literature review? Are there key individuals that you always cite?

L2Q6: Are there challenges specific to PDs?

L1Q6.1: What kind of challenges do you face in publishing (e.g. finding co-authors, journal publication fees, and conference travel expenses)?

L1Q6.2: How do you overcome those challenges?

L2Q7: Do PDs develop their research skills in ways different than NPDs ?

L2Q7.1: Did you take any type of informal education to enhance your research performance?

L2Q7.2: Do you use any tools that support your research publication process?

L2Q7.3: Did you use any of the following approaches for research publication? Please explain: Writing support groups; Structured writing courses; Provision of a writing coach

Appendix B: Stage 2 Survey Questionnaire

1. Full Name

2. Affiliated University

3. Email

4. What is your gender?

   - Male
   - Female

5. What is your marital status?
   - Single
   - Married
   - Separated/divorced
   - Widowed

How many children do you have, if any?

6. What is your last degree?
   - MSc
   - PhD
7. How long did it take you to finish it?
   - 2-3 years
   - 4-5 years
   - More than 5 years

8. From which university did you obtain your last degree?

9. What is your specific field of research?

10. What is the title of your primary current appointment?

- Assistant Lecturer
- Lecturer
- Associate Professor
- Professor or Emeritus Professor

11. Are you the department chair or were you assigned department chair before?

- Yes
- No

If yes, please indicate the start year and end year as department chair

Start year:

End year:

12. For how many of each of the following types of individuals do you currently serve as official advisor?

- Undergraduate Groups (Graduation Projects):
- MSc Students:
- PhD Students:

13. During the past five years, what is your average teaching and administrative work hours per week? (If in your current position for less than five years, base this on the period since your appointment)

14. Have you received any of the following resources during your academic career (Check all that apply)

- Publication financial support
- Research grant
- College scholarship
- Travel funds to attend a conference
- None of the above

15. How many conferences did you attend in the past three years?

16. Please rate the climate of your department based on the following continuum.

17. Are you a graduate of a language school?

- Yes
- No

18. Do you work outside the university?

- Yes
- No

19. How many hours per week do you work outside the university?

20. What motivates you most to publish your research?

- I publish research to get a promotion
- I publish research to stay competitive
- I publish research for international recognition
- I publish research because I enjoy it
- None of the above

21. The majority of your research belongs to which type of the below?

- Studies suggesting new ways of viewing/implementing information processing systems e.g. theories, new architectures, new frameworks, ontologies, network protocols
- Research involving the creation of new information-processing systems
- Research involving the creation and evaluation of tools, formalisms, techniques/methods to support existing information processing systems
- Research on social and economic issues related to information processing systems (Including studies of the social and economic impact of information systems, ethical issues, changing views of humanity, etc.)

22. Which of the below research strategies reflect the majority of your research? (Please check all that apply)

- I prefer to do radical research that suggests new models / frameworks / methods / architecture that weren't implemented before
- I prefer to do incremental research that enhances existing models / frameworks / methods / architectures

- I prefer to map out broad features of important new areas, leaving detailed studies to others
- I prefer to probe deeply and thoroughly in narrow areas
- I prefer research which looks for immediate solutions to real life problems (e.g. social problem or industry need)
- I prefer purely theoretical research
- I prefer to carry out research work pretty much on my own
- I prefer to carry out research within a research team
- I prefer long-term projects to short-term ones
- I prefer short-term projects to long-term ones

23. From where do you get research ideas? (Please check all that apply)

- Publications of researchers I follow on academic platforms (e.g. Google Scholar)
- Live or recorded webinars (e.g. IEEE webinars)
- Papers citing my work
- Conference attendance
- Future work section of papers
- Other (please specify)

24. For each of the below approaches please rate how often do you apply them?

	Never	Seldom	Sometimes	Frequently	Always
Doing research with academics in other universities in Egypt	○	○	○	○	○
Doing research with academics in other departments in my university	○	○	○	○	○
Doing research with academics overseas	○	○	○	○	○

25. Where do you publish your research? (Check all that apply)

- Journals indexed in Scopus
- Journals indexed in Thomson ISI (Clarivate Analytics)
- International Conferences with Proceedings indexed in Scopus
- International Conferences with Proceedings indexed in Thomson ISI (Clarivate Analytics)
- Local Indexed Conferences
- Non-indexed Journals
- Non-indexed Conferences

26. How important are the below factors in determining which journal to publish in?

	Not Important	Slightly Important	Moderately Important	Important	Very Important
The publisher of the journal	○	○	○	○	○
Number of issues per year	○	○	○	○	○
Editorial board	○	○	○	○	○
Journal fees	○	○	○	○	○
Journal impact factor	○	○	○	○	○
Journal SJR (SCImago Journal Rank)	○	○	○	○	○

Other (please specify)

27. How important are the below factors in increasing the chances of acceptance of a paper in a journal/conference?

	Not Important	Slightly Important	Moderately Important	Important	Very Important
Presentation/Structure of the paper	○	○	○	○	○
Reputable co-authors	○	○	○	○	○
Strength of the authors' affiliated universities	○	○	○	○	○
Recency of references	○	○	○	○	○
Including references from the targeted journal/conference proceedings	○	○	○	○	○
Technical depth	○	○	○	○	○
Significance of the contribution	○	○	○	○	○
Theoretical foundation	○	○	○	○	○
Previous publications in the targeted journal/conference	○	○	○	○	○

Other (please specify)

28. Please rate the below publication strategies based on how often you apply them.

	Never	Seldom	Sometimes	Frequently	Always
When I start in a new area of research, I prefer publishing the first paper by myself and then including other authors in the following papers	○	○	○	○	○
I publish part of my research work in a conference before publishing it in a journal	○	○	○	○	○
I submit my paper in top conferences (knowing it might get rejected) before submission in journals to get useful feedback/review	○	○	○	○	○
I submit papers in workshops of top conferences	○	○	○	○	○
I publish papers extending/based on the graduation projects of my last year (undergraduate) students	○	○	○	○	○

29. Please rate the below publication approaches based on how often you apply them.

	Never	Seldom	Sometimes	Frequently	Always
I publish papers with foreign reputable co-authors	○	○	○	○	○
I publish papers in highly ranked journals/conferences	○	○	○	○	○
I publish papers with top publishers (e.g. Elsevier)	○	○	○	○	○
I add my papers in academic networking platforms (e.g. ResearchGate)	○	○	○	○	○
I send hard or soft copies of my paper to researchers in the same field once its published	○	○	○	○	○
I publish papers in specialised journals	○	○	○	○	○
I publish papers in multidisciplinary journals	○	○	○	○	○
I publish papers with new ideas, models or frameworks without experimentation	○	○	○	○	○
I publish papers with new ideas, models or frameworks with experimentation and results	○	○	○	○	○
I publish papers with tools or datasets	○	○	○	○	○
I publish papers in open access journals	○	○	○	○	○

30. What is the primary reason for presenting in conferences?

- Interaction with peers and getting feedback
- To be known among my research community
- To publicise my research and attract paper citation
- To gain knowledge about new research areas and trends
- To search for academic posts, possible grants and project collaborations
- Other (please specify)

31. To what extent are the below research publication challenges applicable on you?

	Not Applicable	Slightly Applicable	Moderately Applicable	Applicable	Very Applicable
Motivation to carry out research is a challenge	○	○	○	○	○
Finding the right journal/conference for my paper is a challenge	○	○	○	○	○
Lack of financial support needed for attending conferences is a challenge	○	○	○	○	○
Proficiency of written English is a challenge	○	○	○	○	○
Formal/Scientific Writing is a challenge	○	○	○	○	○
Time from submission to acceptance in a journal is a challenge	○	○	○	○	○
Insufficient time because of teaching/admin commitments is a challenge	○	○	○	○	○

Other (please specify)

32. Do you use any of the below approaches to overcome these challenges? (Please check all that apply)

- I use journal finder online tools
- I pay for proofreading and editing services for my paper
- I seek external funding agencies (e.g. ITIDA, ASRT, TIEC) to cover the costs of travelling to attend conferences
- I use the financial support provided by the university to cover my travel and publication fees
- I apply for research grants (e.g. Erasmus)
- I establish research teams overseas
- Other (please specify)

33. Do you use any of the below tools in writing and publishing your research? (Please check all that

apply)

- Grammarly
- Reference managers (e.g. Mendeley)

- Latex (e.g. Sharelatex)
- Other (please specify)

34. Do you enhance your publication quality using any of the below approaches? (Please check all that apply)

- Observing highly cited papers to see how they are written and structured
- English writing courses
- Scientific writing / Formal writing courses
- Technical courses related to the field
- Using a graphic designer to represent results in an attractive manner
- Sending papers to friends/relatives for proof editing
- Other (please specify)

35. Do you check the papers of researchers who cited your work

- Yes
- No

36. Why do you track citations mainly?

- To check the geographical distribution of the citing papers
- See the impact of the paper after removing self-citation
- Get ideas on future research areas / improvement areas
- Other (please specify)

37. Do you have an account on any of the below research platforms? (Please check all that apply)

- Academia
- Semantic Scholar
- ResearchGate
- Google Scholar profile
- Arxiv
- DBLP
- ORCID
- ResearcherID
- ACM
- Other (please specify)

38. Please state at least two actions/strategies that you believe could increase paper citation rates.

Action 1:

Action 2:

Chapter Four: Identifying Potential Positive Deviants Across Rice Producing Areas in Indonesia: An Application of Big Data Analytics and Approaches

Basma Albanna, Dharani Dhar Burra and Michael Dyer

Abstract

Approaches to data collection for development programming have commonly relied on official statistical data in combination with surveys to plan, implement and monitor progress and impact for interventions. Today, the emergence of big data has resulted in a paradigm shift, with increasing use of non-traditional data to promote more effective and responsive interventions across various domains. Contributing to the Data Powered Positive Deviance initiative, we conducted data analytics research by merging traditional statistical data with Earth Observation big data to identify and validate potential rice producing villages across Indonesia that might be faring better than others, referred to as positive deviants (PDs). This report details the preliminary results, key learnings, along with some actionable recommendations for future work in the agriculture domain.

Authors' Contributions

BA developed the initial research proposal and the main conceptual idea. BA and DB jointly designed the method and the analytical approach with input from MD. MD forged key partnerships to obtain the administrative data and the Earth Observation data that was used in the analysis. MD and DB scoped, cleaned, pre-processed and prepared Earth Observation and related GIS data for the identification of outliers, as well as prepared the maps that are included in this report. BA sampled and pre-processed the census data and then combined the data set with the Earth Observation data to create the homologous environments and identify the outliers. Inputs on those steps were provided by DB and MD. BA implemented the outlier validation using bivariate analysis; MD implemented the outlier validation using the Google Time Scale tool; and DB implemented the outlier validation using time series Earth Observation data. The results were discussed and interpreted jointly by BA, DB and

MD. BA led the writing of this report in collaboration with DB and MD. All authors jointly identified the limitations of the developed method and provided recommendations for future work.

4.1 Introduction

Public and private organizations working in agriculture development, depend largely on field surveys to identify plot-level management and household-level social, economic, and demographic drivers that determine agricultural productivity. In developing countries, where smallholder agriculture predominates, new and efficient ways of data collection are needed to measure agricultural performance. Current data collection methods fail to capture the complexity of production systems and the varied nature of households that manage these systems, across both spatial and temporal dimensions. For example, traditional field surveys such as the national agricultural census, do not capture certain contributing drivers of productivity, for instance biophysical conditions (e.g. temperature, precipitation, etc.). Nonetheless, earth observation (EO) data has made it possible to map and monitor proxies of croplands and their biophysical environments, which when combined with field surveys and big data analytics, can be used to better characterise the complexity of those production systems. For instance, previous research (Tucker & Sellers 1998; Mkhabela et al. 2011; Bolton & Friedl 2013; Johnson 2014) used the well-established relationship between net primary production and satellite derived measurements of plant phenology such as Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI); which are both measures that are used to quantify vegetation greenness. Those studies demonstrate that the EVI measure is a valid proxy for crop (annuals) yields. And when combined with field surveys, an EVI measure provides an opportunity to better understand the determinants of agricultural productivity, both in terms of controllable factors (i.e. factors farmers can control) and uncontrollable factors (e.g. biophysical variables) captured from EO data. Improved understanding can lead to the identification of over and underperformers (i.e. households/villages with significantly higher and lower agriculture productivity), with relatively higher confidence, from which interventions can be designed to target underperformers using insights gleaned from practices and strategies of the overperformers. Those individuals or groups that are overperformers are referred to as “positive deviants”

(PDs), and adopting their practises and strategies on a wider basis is referred to as the "positive deviance" (PD) approach (Sternin, 2002).

Identifying the overperformers within similar contextual environments was first introduced in 1976 in the case of child stunting to identify dietary practices developed by mothers in low-income families who had well-nourished children (Wishik & Van Der Vynckt 1976). But it wasn't until the early 2000s that PD was promoted as an effective asset based approach for social development after its successful application in rehabilitating an estimated 50,000 malnourished children in 250 communities in Vietnam (Sternin 2002). The PD approach starts by discovering the over performing individuals or communities; following that their underlying practices and strategies are determined; based on which interventions are designed to scale those successful practices from the PDs to the under performers. The underlying assumption is that PDs implement unusual practices and strategies that could provide novel insights to solve complex problems, which conventional solutions fail to solve (Cinner 2016). This type of positive deviance analysis has been also applied in the agricultural domain (Pant 2009; Steinke 2019) and studies show that by analysing what defines a PD within an agricultural community - with "similar" biophysical, socio-economic, and demographic conditions - certain drivers can be repeated or introduced and specific constraints can be removed. Drivers could be external agents, innovative technologies and practices and should require a minimum level of human, social, financial, physical, or natural capital. For instance, in the Brazilian state of Parana, some agricultural communities adopted no-till as a better cultivation method and after demonstrating an increase in productivity, income, and sustainability, the practice was adopted across the state (de Vries 2005).

To the best of our knowledge and according to a recent systematic review on the combined used of PD and big data for development (Albanna & Heeks 2019), previous PD studies focusing on agricultural development, have not combined EO data with administrative data (e.g. agricultural census) to identify PDs and to understand possible drivers of their agricultural performance. In this study, we propose, and trial a method, which combines those two kinds of data, to identify and validate villages (surveyed in the 2013 Indonesian Agricultural Census) that perform substantially better, in terms of agricultural productivity, than their peers despite having similar socio-economic, biophysical and environmental conditions. We used univariate and multivariate outlier detection techniques for PD

identification and used administrative data to understand possible drivers of performance. To validate and denoise the identified PDs, we used the Google earth time scale tool and EO time series data, which further helped us in filtering false PDs from true PDs. Although the scope of this study doesn't fully answer the question "why are some villages faring better than other similar villages" which would require extensive ground surveys, it paves the way for this type of inquiry through providing a spatial targeting method that could significantly reduce the time and cost needed to identify potential PDs in agriculture.

4.2 Data

We relied on two types of data: 1) official administrative datasets i.e. the Agricultural Census Data (2013), and the Village Potential Census data (2014); and 2) EO data that was used to identify homologous environments (i.e. groups of villages that belong to the same biophysical environment). Non-controllable factors, i.e. day temperature and precipitation, that determine agricultural productivity were sourced from Land Surface Temperature and Emissivity data products at 1 km^2 spatial resolution, and monthly temporal resolution, captured since 2004, from NASA's MODIS (Moderate Resolution Imaging Spectroradiometer; Land Surface Temperature and Emissivity (MOD11)), satellite, and CHIRPS (Climate Hazards Group InfraRed Precipitation with Station data), at 0.05 degree arc seconds spatial, and monthly temporal resolution, captured since 1960. In addition, we used earth observation corrected Enhanced Vegetation Index (EVI; derived from MODIS;MODIS/Terra Vegetation Indices Monthly L3 Global 1km SIN Grid (MOD13)), captured at 250 m spatial resolution, and at 16 day time intervals, captured since 2004, as a performance measure, which has been extensively used as a proxy for agriculture performance (Son et al. 2014).

When the research was being conducted, the best available data was relatively outdated (2013 and 2014) but since this work focuses on method development, and subsequent identification of PDs, it was still practical to use this data. In the agriculture census, every house involved in agricultural activity, across Indonesia, is taken as the observational unit. The census contains more than 700 variables related to agriculture production (e.g. type of crops and irrigation systems) per household spanning 18 million agricultural households in Indonesia for each cropping season of the calendar year 2012/2013 for each household. It is important to

note that information is not georeferenced and is collected at the household level and not at the plot level. In order to merge the census data with the EO data, we used the unique village geocode to aggregate all the household data to their respective villages. With rice being a major staple food among Indonesians (Hartini et al. 2005), it is cultivated on significantly larger areas compared to other crops. Therefore, we focused this analysis on rice producing households in the census. In addition, it is relatively easier to estimate productivity of annual crops (such as rice), using proxies derived from EO, as annual crops exhibit "strong" seasonality and clearly distinct temporal features from other land use types. Variables directly related to rice farming practices were selected and aggregated for every village. Specifically, agricultural census data for the third season of the cropping calendar (i.e. between January to April 2013) was used for the analysis. The agriculture census does not capture yields, due to its differential design and end use, therefore, to circumvent this issue, we used EVI, aggregated to the village as the performance measure.

In parallel, to include socio-economic and demographic information about the households, we used the 2014 village potential survey (PODES). The survey is collected by a different directorate (from the directorate that conducts the Agricultural Census) in the Ministry of National Development Planning (BAPPENAS) by interviewing the head of each village in Indonesia. The 2014 PODES data had more than 800 variables relating to village characteristics such as water resources, public services and facilities, market assets, etc. Similar to the agricultural census data, the PODES data is not geotagged. We selected variables of potential relevance (selected variables can be found in the Appendix) from the agricultural census and PODES, and aggregated the agricultural census data to the village level (a mode function was used for categorical variables and a proportion or average function was used for numerical variables), and joined both datasets with the unique geocode for the village. Notably, selection of variables is a context specific activity, and was dependent on the research question.

Since yields were not captured in the census data, we relied on a commonly used EO metric - the corrected Enhanced Vegetation Index (EVI) - as a proxy for agricultural performance, because it is a well-known measure of plant greenness or leaf area index (Son et al. 2014) and was developed to optimize vegetation signals in regions with high biomass and has less saturation from when compared to the NDVI (Huete et al. 2002; Qiu et al. 2013). To derive

the EVI, we used MODIS satellite imagery, which has global coverage and a 250 m spatial resolution, with a 16-day revisiting interval. Across the agricultural season between January and April 2013, a global EVI 250m pixel resolution raster layer, was clipped to the extent of Indonesia and was extracted in the Mercator projection, for each time point. Across the raster brick, the Maximum value for each pixel, across all time-points, was extracted. Since we aggregated the agriculture census and PODES data to the village, the Maximum EVI value for each pixel was also aggregated to the village.

EVI values are sensitive to crop types and the obtained Maximum EVI values could reflect other crop types that are grown in the village. To control for this source of error, we extracted the average Maximum EVI values for the rice growing areas within village boundaries with a rice crop mask provided by the Indonesian Ministry of Forestry. For the purpose of aggregation, Maximum EVI values for each pixel were averaged across all pixels that belong to a village. Village boundary data in a shapefile format were provided by the Indonesian Bureau of National Statistics.

Monthly rainfall data for the season between January 2013 and April 2013, was obtained as raster layers (0.05 arc seconds) from CHIRPS. This is an open source globally gridded dataset, containing 35 years of rainfall data, produced and maintained by the University of California San Diego and USAID. CHIRPS is a hybrid data product, that grids global weather station data, and interpolates the weather station data, with satellite-based precipitation estimates, obtained from NASA's Global Precipitation Missions (GPM) and NOAA's CPC merged analysis of precipitation (CMAP). Although this smart data-fusion approach removes systematic bias associated with the weather station data, the CHIRPS dataset still suffers from issues such as underestimated precipitation measures in complex orography. For this analysis, monthly temperature data for the selected cropping season was downloaded as individual raster files. For each month, the rasters were merged with Indonesia's official administrative boundary shapefile, and values of pixels belonging to each village were averaged separately for each month, to obtain monthly average rainfall (in millimeters) for each village.

The monthly temperature data from Land Surface Temperature and Emissivity data product of MODIS, at 1km spatial resolution, was downloaded as raster files, separately for each

month within the selected cropping cycle. To obtain average monthly temperature values, the raster files were merged with the administrative boundary shapefile, and the pixel values averaged across, and extracted at the village level. The temperature data from the MODIS sensor was obtained in digital numbers (DN), which were then converted to temperature in degrees centigrade, by multiplying the DN value with 0.02 (i.e. scale factor), to obtain temperature in Kelvin, and then subtracting with 273.15 to obtain temperature values in degree centigrade.

In summary, the following datasets were used for the analysis:

- Average monthly rainfall data (in millimeters for the cropping season between January to April 2013) from Climate Hazards Group InfraRed Precipitation with station data (CHIRPS),
- Average monthly temperature data (in degree centigrade, for the cropping season between January to April 2013) from the Land Surface Temperature and Emissivity data product of MODIS
- Averaged Maximum EVI for each village obtained from bi-weekly MODIS EVI data produced every 16 days for the cropping season between January to April 2013
- Land use 2014 rice crop mask data intersected with the village boundaries in order to extract EVI of the rice areas in each village
- 2013 Indonesia Agricultural Census Data capturing agriculture production data for more than 18 million households that are involved in agriculture for the seasons between years 2012 and 2013
- 2014 Indonesia Village Potential Data for 82,000 villages
- 2014 administrative boundaries of villages data (bureau of statistics)

The following datasets were used for subsequent validation of the results:

- Monthly aggregates of precipitation (in millimeters) from January 2001, until December 2015, for Indonesia at 5 square kilometre (0.5 arc seconds) resolution, from Climate Hazards Group InfraRed Precipitation with station data (CHIRPS)

- Monthly aggregates of daytime temperatures (in degree centigrade) from January 2001, until December 2015, for Indonesia at 4 square kilometre resolution, from the MOD11 data product of MODIS
- Monthly aggregates of EVI from January 2001, until December 2015, for Indonesia at 1 square kilometre resolution, from the MOD13 data product of MODIS

Figure 12 presents the rice growing area in Indonesia in the year 2014. It is important to note that the rice areas might be slightly different in 2014 than what is stated in the 2013 census data. We used the rice mask to reduce the error of capturing EVI values from non-rice areas instead of using if we captured EVIs for the entire village. Seen in Figure 12, this rice mask covers both types of rice, the wetland and the dryland, without differentiating between the two. Therefore, our sample included villages growing both types of rice.

Figure 12: Rice crop mask in Indonesia - 2014

4.3 Study Sample

Our data sample included villages in Indonesia having at least one household growing any type of rice (i.e. dryland or wetland rice) in the third season (i.e. January to April 2013). Combined with average monthly rainfall and temperature, and average Maximum EVI during the cropping season between January to April 2013. According to the agricultural census, there are 41,664 villages growing rice in Indonesia. However, we were able to extract EVI values for only 18,978 villages. To prevent cross-signal issues from other crops, we used a rice mask layer

for each village, to extract average Maximum EVI values. However, the rice mask was recorded for the year 2014 and had limited metadata about which cropping season and rice variation it represented. The significant reduction in the number of villages in the census, for which EVI values could be obtained, can be attributed to this temporal mismatch between the rice crop mask layer and the agricultural census data. It is also possible that the rice mask for 2014 represented a different cropping season than the cropping cycle selected for the analysis, or there could be a temporal shift in the amount of rice production in 2014, compared to 2013. The total of 18,978 villages was further reduced to 17,517 villages, due to missing temperature and/or rainfall data or not being captured in the PODES census data. Our final data sample of 17,517 villages constituted a total of 4,051,416 households growing rice, wherein each village had data from the agricultural census, PODES, average monthly rainfall and temperature data, and an average Maximum EVI, all for the cropping cycle between January and April 2013.

4.4 Methodology

The primary objective of this study was to develop a method that combines EO data with existing, varied administrative data, to identify rice villages that are performing significantly better than rice villages having similar conditions. To test the viability of the proposed method it is important to validate the identified PDs by checking whether they are true PDs or not. Figure 13 provides a summary of the approaches used for potential PD identification, which are explained in further details in the sections to follow. In this study, all the data analysis was done using the statistical software R v3.4.1, and QGIS v3.8 and ArcGIS v10.7.1 for spatial processing.

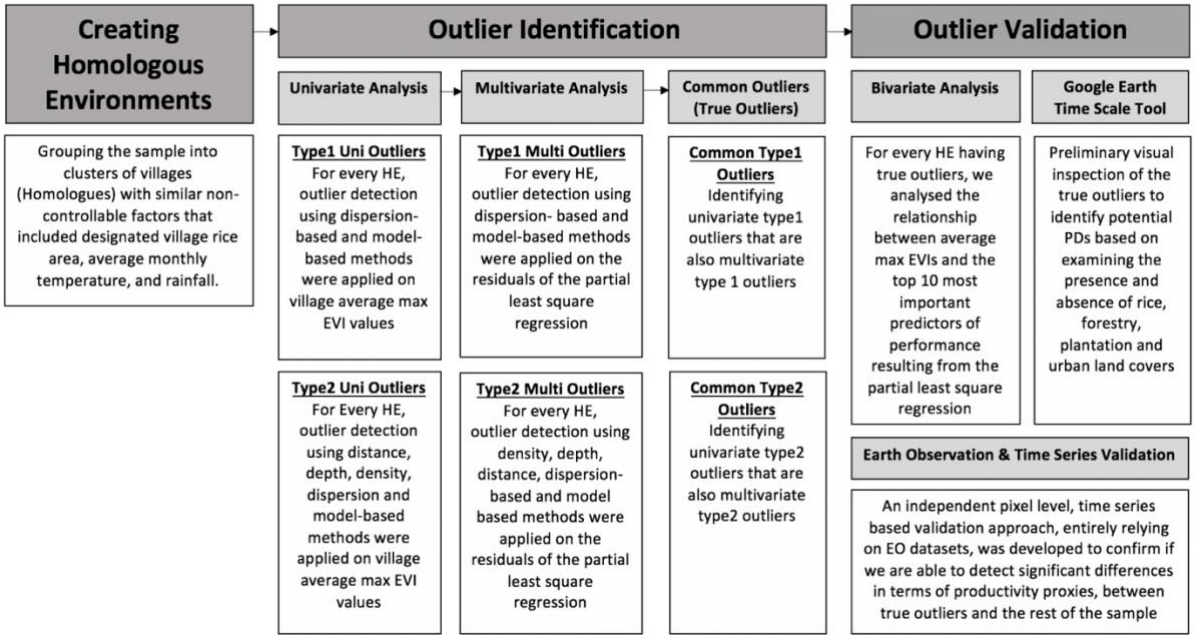

Figure 13: A summary of the approaches used for potential PD identification and validation

4.4.1 Creating Homologous Environments

The first step of this methodology was to create homologous environments (HEs) i.e. grouping the 17,517 villages, into clusters of villages with similar non-controllable factors, that included designated village rice area, average monthly temperature (in Degree Celsius) and rainfall (in millimetres). The total rice area was clustered using the *densityMclust* function of the mclust R Package into 5 clusters, such that each cluster would have groups of villages with similar total rice areas. The number of clusters (k=7) was determined automatically based on the score of the Bayesian Information Criterion (BIC). The number of villages and the mean total rice area in every cluster is shown in Table 19.

Area Clusters	Number of Villages	Mean of Village total rice area (meter square)
Cluster 1	815	4,231
Cluster 2	1,881	49,495
Cluster 3	6,158	2,811,223
Cluster 4	6,459	1,108,841
Cluster 5	2,204	3,922,707

Table 19: Village clusters based on rice area

Next, when clustering villages based on biophysical environments (i.e. based on average temperature and rainfall), we performed hierarchical clustering with principal components using the *HCPC* function in the factominer R Package. This function performs an agglomerative hierarchical clustering on results from a factor analysis. The first step was to perform a principal component analysis (PCA) on a dataset containing 8 columns of normalized. PCA looks for a few linear combinations called "Principal components" (PCs) that can be used to summarize the data without losing too much information (Maitra & Yan 2008). Ordered quantile normalization was done using the *orderNorm* function in the bestNormalize R package We chose to retain the three principal components that were able to explain 80% of the variance using the "elbow" method applied on a scree plot. The output of the PCA was then passed to the *HCPC* function which clustered the villages into three biophysical clusters containing 7135, 5348, and 5034 villages respectively. Based on the area clusters and biophysical environment clusters, we developed 15 HEs, wherein each homologous environment consisted of villages having similar rice area and biophysical conditions as shown in table 20.

Homologue	Biophysical Cluster	Area Cluster	Number of Villages
11	1	1	259
12	1	2	667
13	1	3	2,532
14	1	4	2,943
15	1	5	734
21	2	1	163
22	2	2	436
23	2	3	1,361
24	2	4	2,438
25	2	5	950
31	3	1	393
32	3	2	778
33	3	3	2,265
34	3	4	1,078
35	3	5	520

Table 20: Number of villages in each homologue

4.4.2 Outlier Identification

Before identification of potential PDs in each HE, we used two different, yet complementary approaches, to identify outliers within each HE. The two approaches are as follows:

- Univariate analysis, where outlier villages in each HE are identified based on the performance measure, i.e. average Max EVI value for each village
- Multivariate analysis, where outlier villages are identified in a relative sense using multiple control variables (the selected controllable factors from Agricultural census and PODES), that could determine average Max EVI values for each village

The final set of outlier villages would be those that consistently appear in both types of analysis. The reason we perform two types of analysis was to ensure that outliers are identified with higher confidence, after controlling for multiple factors, and are not merely identified by chance. The universe of factors, used as controls in this study, are derived from a comprehensive list of factors, that are collected repeatedly, across time by governments. If a village continues to be an outlier, with (multivariate analysis) and without (univariate analysis) the use of these controllable factors, it would suggest, that practices in the outlier villages are different from the rest of the villages, within the same HE, and are not completely captured by the current universe of factors, which further increases the chances of identifying potential PDs, from this list of outliers, in the later stages of the analysis.

Univariate Analysis

In this analysis, for every HE, the average Max EVI values, for all the villages, were used to identify outliers. As expected, the average Max EVI values of villages within an HE didn't follow a normal distribution, further indicating the presence of complex, non-uniform production systems within HE. Therefore, a method that identifies observations as outliers, solely based on the assumption of normal distribution cannot be used. Instead we used an outlier detection approach, that uses multiple measures, in addition to the distribution of the performance measure, to identify outliers. We used the OutlierDetection function in the OutlierDetection R Package, which identifies outliers using a combination of different

methods. Since the focus is to identify over-performing villages, we term them as positive outliers. Two types of outliers were identified using this function:

- **Type1:** Outliers are identified using the default OutlierDetection function which finds outlier observations using dispersion based and model based methods for outlier detection. The total number of positive outliers, across all HEs, identified using this method is 144 villages.
- **Type2**: Outliers are identified using the OutlierDetection function but after adding depth, density and distance methods to the dispersion-based and model-based methods for outlier identification. It can be considered as a narrower filter for outlier detection, hence, it resulted in a smaller number of outliers. The total number of outliers identified using this method is 29 villages.

Table 21 presents the distribution of Type1 and Type2 outliers across the 15 homologues. The above approach was unable to identify outliers in certain HEs (e.g. "11", "21", "22" and "31"). Alternatively, this approach yielded only Type1 outliers and not Type2, in certain HEs (e.g. 12 and 35).

Homologue	Number of Villages	Type1 outliers	Type2 outliers
11	259	0	0
12	667	4	0
13	2,532	27	2
14	2,943	22	4
15	734	4	2
21	163	0	0
22	436	0	0
23	1,361	1	1
24	2,438	20	1
25	950	15	5
31	393	0	0
32	778	2	2
33	2,265	30	7
34	1,078	11	5
35	520	7	0
Total	**17,517**	**144**	**29**

Table 21: Number of PDs in each homologue

Multivariate Analysis

In the previous approach, we used only the average Max EVI value, to identify positive outliers, but we didn't consider other drivers of agricultural performance that could have influenced those values. Those drivers include but are not limited to the village income, crop ecosystems, other plantation farming activities and possible environmental stresses. In this section, we present results from the multivariate analysis that was applied to identify positive outlier villages having an observed Max EVI value that is significantly higher than the predicted Max EVI, which was modelled based on possible drivers of performance. Variable selection and dimensionality reduction is a crucial step in multivariate analysis, especially when you have a large number of possible explanatory/predictive variables (our data sample had 75 variables). Additionally, if the independent variables are highly correlated, they increase the variance in the regression estimates, and this requires special methods of regression that could overcome this problem of multicollinearity (Kleinbaum et al. 1988). Among those methods, is the PCA and Partial Least Square (PLS). In principal component regression, the PCs are used to predict the dependent variable, which in our case is the average Max EVI.

One drawback of doing regressions using PCA is that the selection of the principal components doesn't give much importance to how each independent variable may be related to the dependent variable, as it is an unsupervised dimensionality reduction technique. Since we are trying to capture possible drivers of EVI, it is crucial to reduce the dimensionality of the data by identifying PCs that not only summarize the independent/predictor variables, but that are also related to the dependent/outcome variable. PLS allows us to achieve this balance by using a dimensionality reduction technique that is supervised by the outcome variable (Maitra & Yan 2008). In comparison to PCA, PLS regression achieved lower RMSE, higher R-square scores and higher percentage of explained variance in the outcome variable. The r function train in the caret R Package was used to compute the PLS regression by invoking the pls R Package. The numeric variables were scaled using the scale function in the base Package to make them comparable with the categorical variables which were transformed into dummy variables in the PLS regression. Cross validation was used to identify the optimal number of PCs to be incorporated in the model. The optimal number of components is the

one that achieves the lowest cross validation error (RMSE). For each of the 15 HEs, PLS regression was applied, and the optimum number of PCs were used to model the predictor variable. The PC residuals (i.e. the difference between the observed value and the fitted value of the outcome variable predicted by the PLS principle components) were used for outlier analysis using the OutlierDetection function in the OutlierDetection Package. In a similar way as the univariate analysis, two types of PDs were identified in the multivariate analysis. The first type used the default methods and the second type used a combination of all methods (i.e. density, depth, dispersion and distance) for outlier identification. While the univariate and multivariate outlier analyses detected 144 and 539 Type1 outlier villages respectively, only 32 Type1 outlier villages were common to both sets of analyses. Similarly, while the univariate and multivariate outlier analyses detected 29 and 48 Type2 outlier villages respectively, only 6 Type2 outlier villages were common to both sets of analyses.

Table 22 summarizes the number of components used in modelling the outcome variable in the PLS regression which was applied for each HE separately. It also presents the cumulative proportion of variance explained, the RMSE and r square scores, the positive outliers identified by multivariate analysis and the common outliers which were also identified using the univariate analysis for each of the two types of outlier detection. In total, out of the 15 HEs, there were nine HEs that had common Type1 outliers and three HEs that had common Type2 outliers. Common here refers to outliers identified by both univariate and multivariate approaches. These common outliers, specifically for Type1 and Type2, are now referred to as True Outliers. It is also interesting to see that for few HEs, all outliers identified using the univariate analysis remained as outliers in the multivariate analysis too. For example, in homologue 23, there was a univariate Type1 outlier village that is also a multivariate Type1 outlier village.

HE ID	Max EVI % of explained variance	RMSE	R	Type 1 Outliers			Type 2 Outliers		
				Uni	Multi	Common (True Outliers)	Uni	Multi	Common (True Outliers)
11	27.3%	0.82	0.07	0	1	0	0	1	0

12	41.4%	0.81	0.35	4	12	0	0	1	0
13	39.3%	0.80	0.32	27	136	5	2	5	0
14	33.5%	0.91	0.25	22	79	4	4	8	2
15	44.7%	0.73	0.45	4	17	2	2	2	1
21	79.9%	0.69	0.61	0	5	0	0	1	0
22	74.9%	0.62	0.63	0	14	0	0	3	0
23	49.1%	0.69	0.54	1	20	1	1	3	0
24	31.6%	0.83	0.28	20	69	0	1	3	0
25	34.4.%	0.90	0.22	15	23	1	5	1	0
31	41.8%	0.80	0.23	0	8	0	0	1	0
32	43.7%	0.72	0.44	2	29	1	2	5	0
33	30.2%	0.85	0.23	30	80	13	7	8	3
34	35.1%	0.80	0.29	11	35	3	5	4	0
35	47.7%	0.97	0.12	7	11	2	0	2	0
Total number of PDs				144	539	32	29	48	6

Table 22: PLS regression Statistics

The "**Max EVI % of variance explained**" column in Table 22, suggests that despite controlling for various factors, there are several controllable and uncontrollable factors, which are not captured by administrative data, that contribute to the variance of observed performance between villages within a HE. The *varImp* function in the caret R package was used to identify the most important predictor variables (i.e. controllable factors) in the model produced by the *train* function. For PLS regression, the variable importance measure is based on the weighted sum of the absolute regression coefficients. The weights are a function of the reduction of the sums of squares, across the number of PLS components and are computed separately for each outcome. Therefore, the contribution of the coefficients is weighted

proportionally based on its ability to reduce the sums of squares. The top 10 important variables in the nine HEs containing common outliers (outliers that were identified by both the multivariate and univariate analysis) were analysed again, to identify variables that were common across HEs and variables that were specific to them. In total 35 variables collectively appeared in the top 10 list in each of the nine homologues. Table 23 provides a summary of those variables and how they are ordered across the different homologues.

Important Variables		**Homologue**								
Name	Code	13	14	15	23	25	32	33	34	35
Doing plantation farming	r2042	1	1	-	2	-	3	1	1	7
Age of main farmer	r216	2	2	-	3	-	4	9	6	5
% of households growing dryland rice in Season3	r301bk4	6	-	6	-	9	1	4	-	9
% of rain fed irrigation	r901a3k2	-	4	3	-	10	6	8	3	-
% of households growing wetland rice in Season3	r301ak4	7	-	10	-	8	5	5	-	8
Flood Events in 2013	R601B_K7	10	-	-	4	1	7	-	-	10
Number of families without electricity	R501B	8	7	-	-	-	-	3	5	3
% of households growing dryland rice in Season2	r301ak2	-	-	5	9	-	8	-	-	1
Number of markets without buildings	R1205	-	-	-	1	-	2	-	2	4
% of households growing wetland rice in Season2	r301ak3	5	8	4	-	-	9	-	-	-
Cooking Fuel used is "LPG"	R5032	-	9	-	-	-	-	-	-	2
Number of active saving and loan cooperatives	R1212C	9	-	-	-	-	-	2	4	-
Main cooking fuel used is firewood	R5034	-	6	-	-	-	-	-	-	-
Village Revenue	R1501A_K3	-	-	1	-	4	-	-	-	-
Main source of income for the majority of the population is plantation	R404B14	3	-	-	-	-	-	6	8	-
Number of female migrant workers	R403B2	-	-	-	-	3	-	-	-	6

Number of landslides	R601A_K7	-	-	-	-	-	-	10	-	-
% of technical irrigation	r901a1k2	-	5	-	-	-	-	-	-	-
Main type of household business is plantation	r214204	4	-	-	-	-	-	7	-	-
Proportion of simple irrigation	r901a2k2	-	-	2	-	-	-	-	-	-
Doing Horticulture Activities	r2032	-	10	-	-	-	-	-	-	-
Majority of wetland rice is managed with revenue sharing	r301ak82	-	-	-	-	5	10	-	-	-
Drainage through river/irrigation channel/lake/sea	R5064	-	-	-	10	-	-	-	-	-
Water source for bathing is well	R507B4	-	-	8	6	-	-	-	-	-
Burning of fields before farming	R5132	-	-	-	-	-	-	-	7	-
Number of male migrant workers	R403B1	-	-	9	-	2	-	-	-	-
Doing Aquaculture Activities	r2082	-	3	-	-	-	-	-	-	-
Water source for bathing is drilling well or pump	R507B3	-	-	-	8	-	-	-	-	-
Drainage through sewage system	R5062	-	-	-	-	7	-	-	-	-
Road surface type from production centre to the main village road is land	R404B23	-	-	-	7	-	-	-	-	-
Village area that borders by the sea	R307A2	-	-	-	-	-	-	-	9	-
Pollution Incidents	R512A_K22	-	-	-	-	6	-	-	-	-
The existence of settlements	R511A2	-	-	-	5	-	-	-	-	-
Water source of bathing is river/lake or pond	R507B6	-	-	7	-	-	-	-	-	-
Utilization of the sea for public transportation	R307B1E2	-	-	-	-	-	-	-	10	-

Table 23: PLS important variables

As shown in table 23, doing plantation farming along with rice was identified as a key predictor of average Max EVI values. Villages with the majority of households doing plantation farming were associated with better average Max EVI values. It ranked first in four out of the nine analysed HEs, and on average, it is the most highly ranked variable. The age of the main farmer came second, it appeared as one of the top 10 variables in 7 out of the 9 HEs. As the average age of the main farmer in a village increases, the average Max EVI value decreases. Other top predictors included the proportion of rice households with rain fed irrigation (as it increases, the average Max EVI values increases), the number of flood events (as it increases, the average Max EVI values decreases) and the existence of families without electricity (as it increases, the average Max EVI values increases). There were also predictors that were specific to certain homologues, like aquaculture household activities and horticulture farming in HE "14", the existence of settlements in HE "23", burning of the field in preparation of the agricultural land in HE "34" and pollution incidents in HE "24".

4.4.3 Outlier Validation

The previous steps, i.e. construction of HEs and identification of true outliers, rely on several assumptions, and are performed at a higher aggregation level. For instance, the EO data sources used in the previous steps, provide data at a very high spatial resolution (e.g .precipitation data from CHIRPS is provided at 5 square kilometre resolution), or at a higher temporal resolution (e.g. EVI values are generated once every 16 days). In order to merge these EO data sources with administrative data (such as the agricultural census), that come at a different spatial and temporal resolution, the EO data is aggregated to the smallest administrative unit, i.e. the village level, at which the agricultural census is often collected. In fact, since the agriculture census is collected at the household level, we also aggregate every variable to the village level. Additionally, since administrative data is collected at large scales, systematic errors especially during data collection and (pre) processing, can occur. Therefore, it is necessary to validate whether the identified true outliers are potential PDs. Validation of true outliers needs to happen in terms of errors resulting due to: 1) aggregation of the various data sources used; 2) systematic errors during data collection/processing of the administrative data and 3) False positives resulting from the use of varied statistical approaches such as clustering (to identify HEs), outlier detection and PLS regression. To

specifically address potential errors arising from the above mentioned sources, three outlier validation approaches were conducted:

- **Bivariate analysis:** For every HE containing common true outliers, bivariate analysis was conducted to understand the relationship between each of the top 10 variables (resulting from the PLS regression) and the average Max EVI. The bivariate analysis explained in the next section, describes the contribution of each variable, in explaining the observed variance (which also includes the additional variance), on the average Max EVI, and how this differs across HEs. Validation here was done through 1) identifying if there are variables that are common across true outliers that are known (based on the literature) to have a positive impact on agriculture productivity and 2) if growing rice is the main driver of average Max EVI for true outliers, and it is not plantation what is driving their EVIs.
- **Google Earth time scale tool:** A preliminary visual inspection to assess coherence between the results of PLS/bivariate analysis, performed to identify the true outliers, and from generalized assessments, resulting from manual labelling and examination of a subset of factors, identified in the PLS/bivariate analysis, using the time scale tool in Google Earth Pro.
- **Earth observation and time series analysis:** An independent pixel level, time series based validation approach, entirely relying on EO datasets, was developed to confirm if we are able to detect significant differences in terms of productivity proxies, between outliers and the rest of the sample. To build, and test this method, HE21 and 22, that had the highest EVI variance explanation in the PLS were used.

Bivariate Analysis

HE "13": This HE contains 2532 villages out of which five villages are true (common) outliers. As shown in figure 14, plantation farming was one of the key predictors of average Max EVI and four out of the five PD villages did plantation farming along with rice farming, three of which had plantation farming as the main type of household business and two had plantation farming as the main source of income. The average age of the main farmer was also identified as one of the top predictors of Max EVI. As the average age increases, the Max EVI decreases. The majority of true outlier villages had an average main farmer age around the thirties. The

figure also shows that as the proportion of households growing dryland rice in season three increases, the average Max EVI increases, but this is not the case with the proportion of wetland rice, which affects Max average EVI negatively. The true outliers were divided between both groups, two villages had the majority of their households growing wetland, and two had the majority of households growing dryland rice. And the wetland rice true outliers didn't grow rice in season two. Additionally, the existence of active saving and loan cooperatives was inversely proportional to average Max EVI values, as true outliers were villages having zero cooperatives. Figure 14 also shows that as the number of flood events increases, average Max EVI decreases and true outlier villages had 0 to 3 floods in the year 2013.

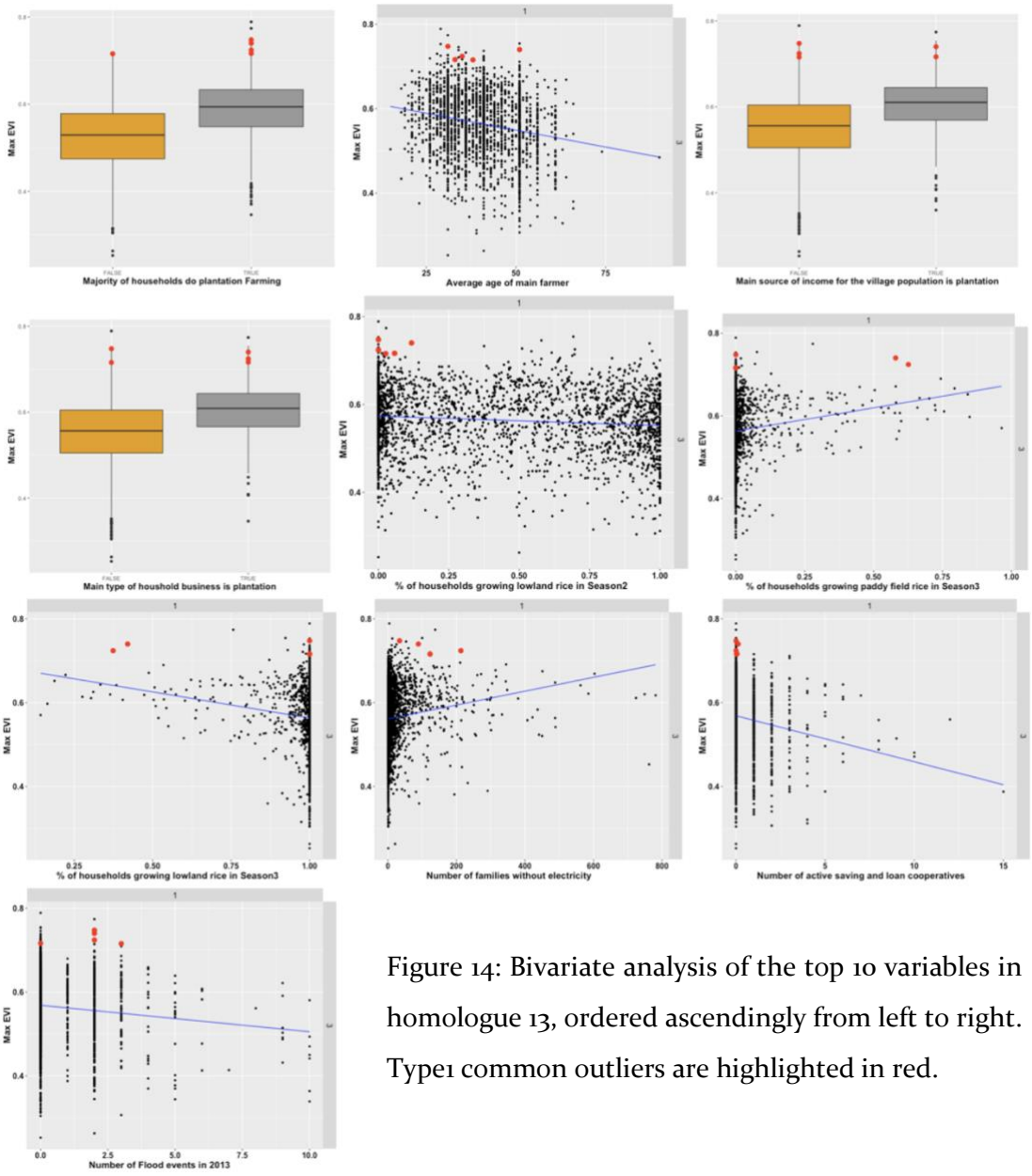

Figure 14: Bivariate analysis of the top 10 variables in homologue 13, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "14": contains 2943 villages out of which four villages are common PDs. As shown in figure 15, doing plantation farming was also one of the key predictors of Max EVI, but three of the four identified true outliers didn't do plantation farming, despite its association with higher average Max EVI values. The average age of the main farmer is also one of the most important predictors, and true outlier villages had an average main farmer age ranging from 25 to 50 years old. In this homologue, doing aquaculture household activities and horticulture

farming were identified among the top predictors of the outcome variable and they were associated with higher average Max EVI values, however none of the true outliers did aquaculture household activities and one outlier did horticulture farming activities. Figure 15 also shows that rain fed and technical irrigation were identified as top predictors of average Max EVI, the higher the proportion of the former the better the EVI values while the latter showed a slight decrease in Max EVI values when there is an increase in technical irrigation. As the number of families without electricity increases (i.e. rural areas) the EVI values increases, however true outliers were located at the lower end with zero to few families without electricity. Similar to homologue "13", as the proportion of households growing wetland rice in season 2 increases, Max EVI values decreases and true outliers were evenly found at the two ends of the spectrum.

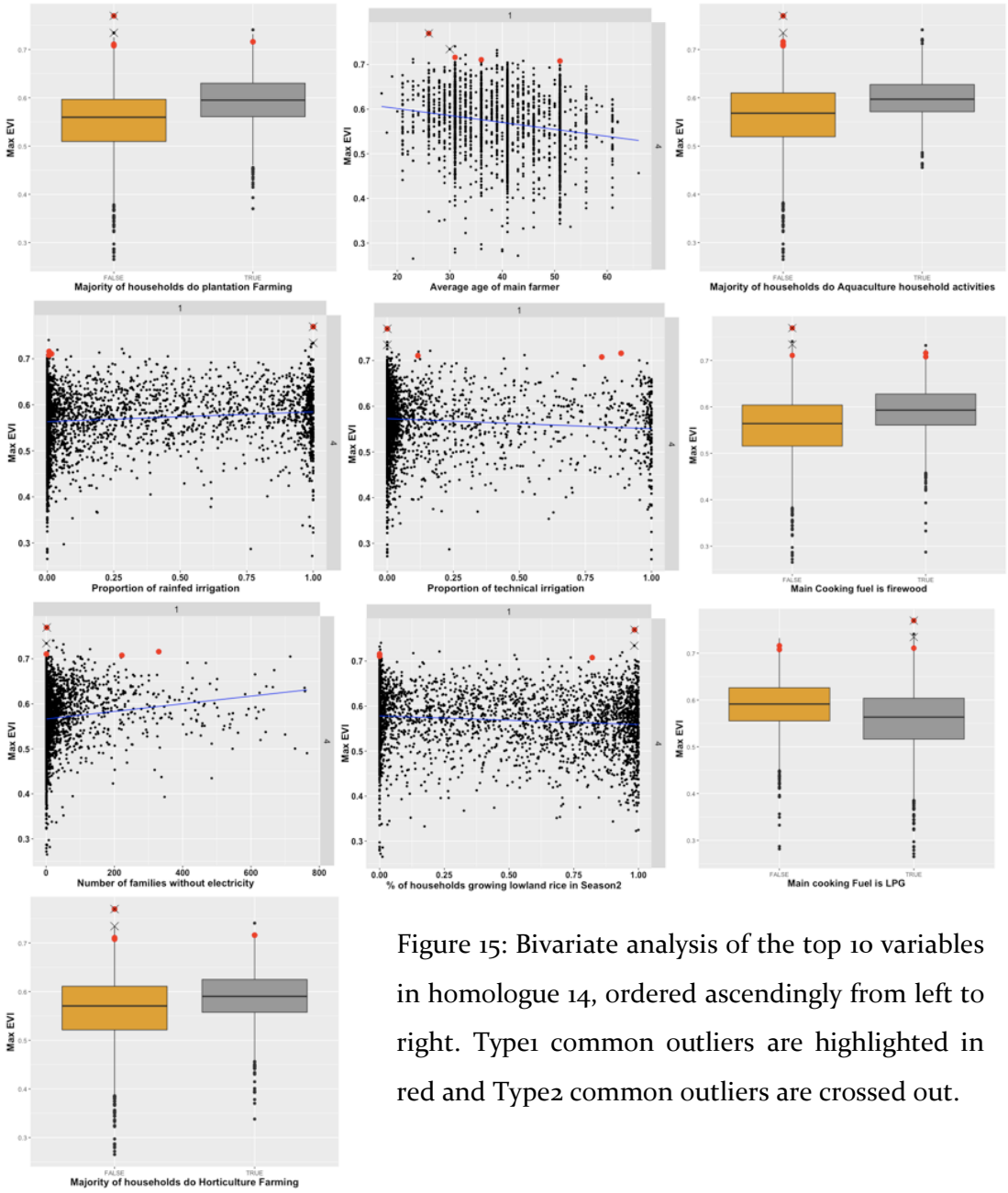

Figure 15: Bivariate analysis of the top 10 variables in homologue 14, ordered ascendingly from left to right. Type1 common outliers are highlighted in red and Type2 common outliers are crossed out.

HE "15": contains 734 villages out of which 2 villages are common true outliers. As shown in figure 16, the most important predictor is village revenue, as it increases Max EVI values increases. One outlier was at the lower end (revenue for this village wasn't provided) and the other outlier was at the higher end. Figure 4 also shows that rain fed and simple irrigation were identified as top predictors of Max EVI, the higher the proportion of households with them the better the EVI values. One outlier village had all households with rain fed irrigation

and the other outlier had almost half the households with simple irrigation and the other half with rain fed irrigation. None of the outliers grew dryland rice in season three although it's directly proportional with Max EVI and despite the fact that both PDs had respectively around 0.25% and 0.95% of their households growing dryland rice in season two. However, both PDs had households growing wetland rice as well with similar percentages in season two. The main water sources for bathing in the village was also identified as an important predictor. Rivers, lakes or ponds were associated with lower Max EVIs and using wells for bathing were associated with higher Max EVIs. However, true outliers were divided evenly across both sources. Figure 16 also shows that as the number of male migrant workers increases, EVI values increase. However, both outliers had a very small number of migrant workers. Almost all households in outlier villages grew wetland rice in season three despite the association with lower Max EVI values.

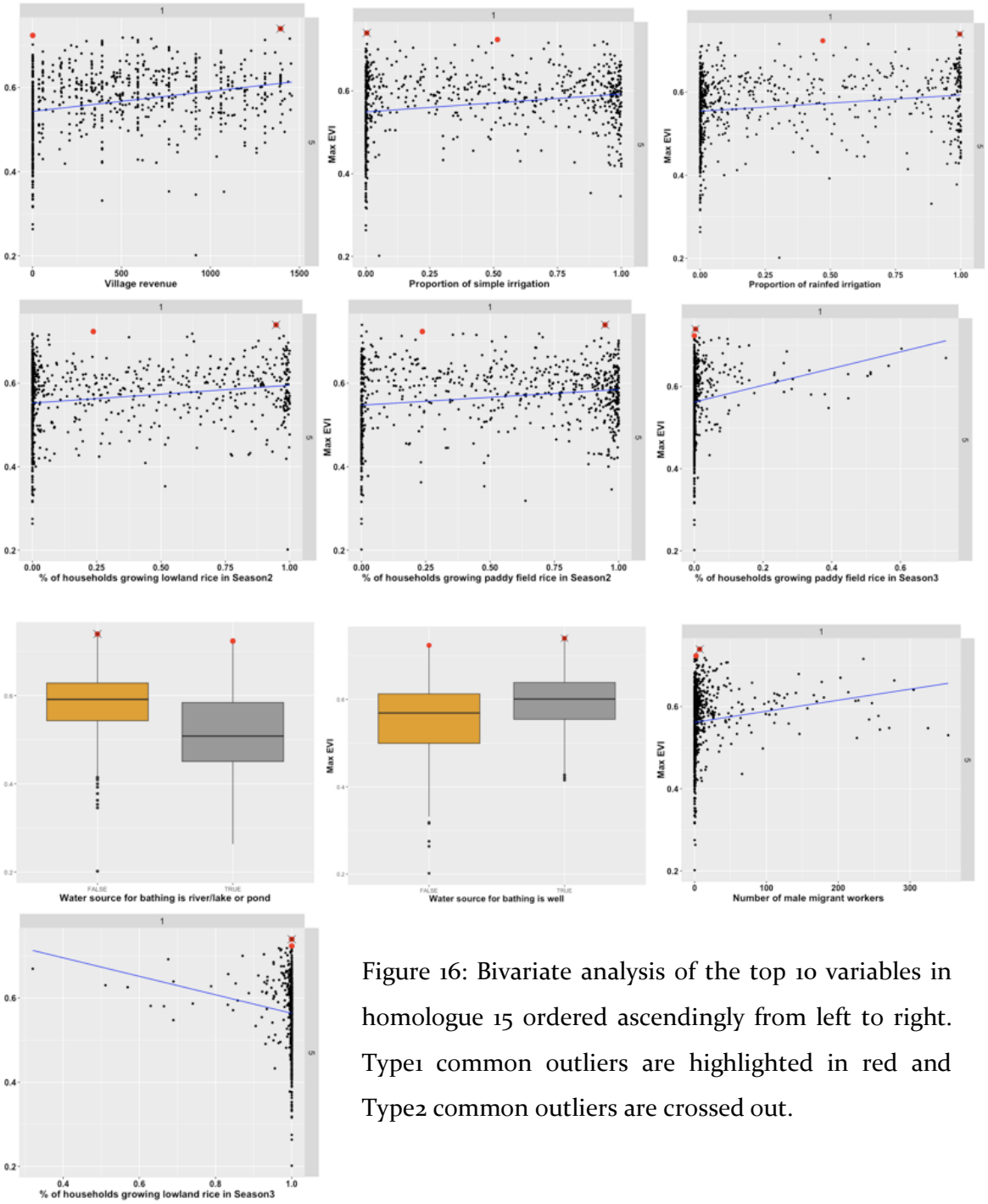

Figure 16: Bivariate analysis of the top 10 variables in homologue 15 ordered ascendingly from left to right. Type1 common outliers are highlighted in red and Type2 common outliers are crossed out.

HE "23": contains 1361 villages out of which only one village is a true outlier. As shown in figure 17, the most important predictor of Max EVI is the number of markets without buildings. As the number of markets without buildings increases, the Max EVI values decreases and the outlier village had no such markets. Doing plantation farming appeared

again as an important predictor and the outlier village did plantation farming. Average age of the main farmer appeared again as an important predictor and the outlier village average age of the main farmer was in the forties. Figure 17 also shows that the number of flood events have a negative impact on the outcome variable and the outlier village had no flood events in the year 2013. The results also show the existence of settlements and the use of water pumps for bathing is inversely proportional to Max EVIs (the outlier village had non) and using wells as the primary source of water for bathing is directly proportional to Max EVIs (the outlier village used it). The existence of a land road surface from the production centre to the main village road was associated with better EVI values and the outlier village had a land road surface. Drainage through irrigation channels, lakes or in the seas was associated with higher Max EVI values and the outlier village didn't have such drainage systems. Finally, the outlier village had almost zero households growing dryland rice in season two.

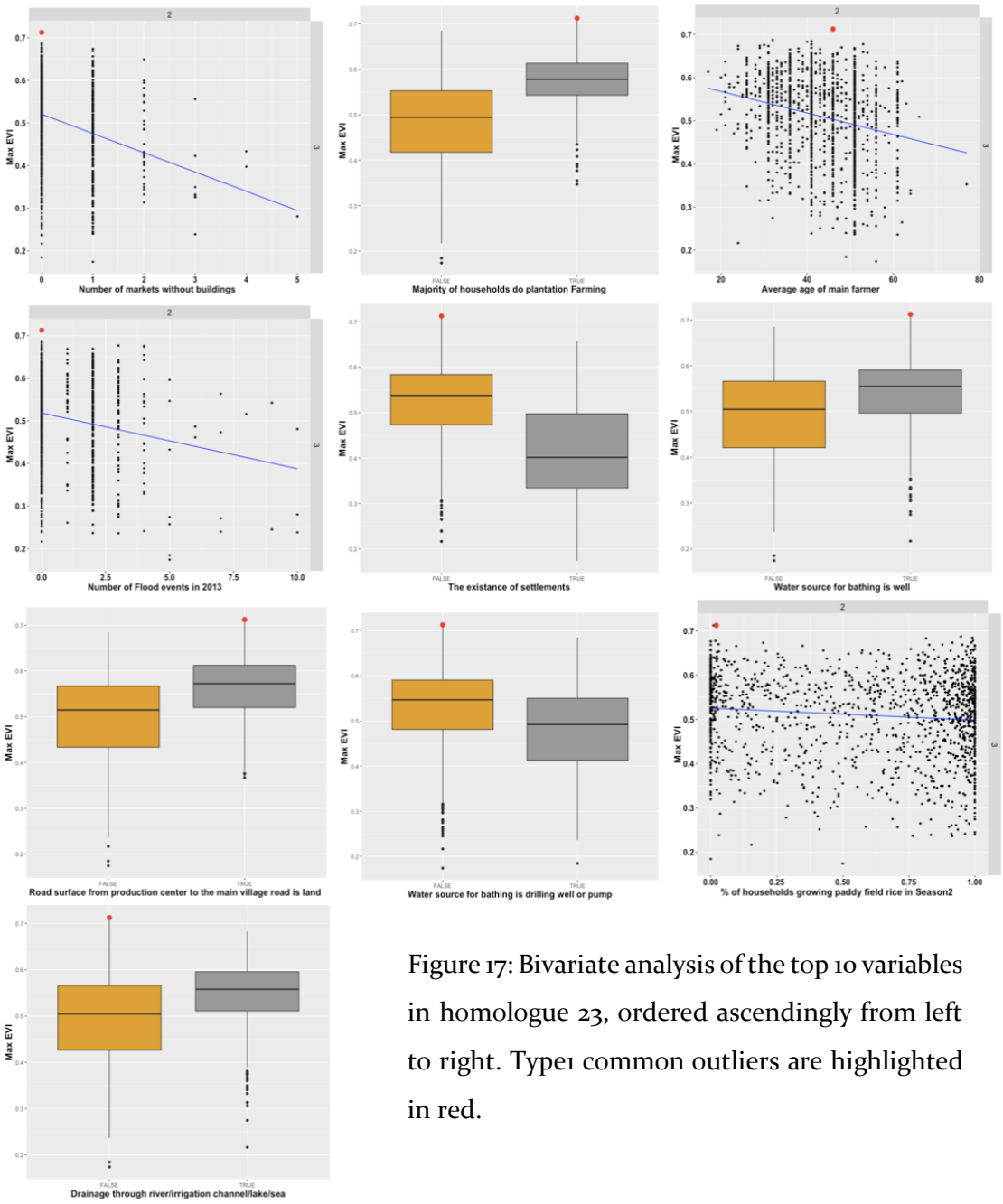

Figure 17: Bivariate analysis of the top 10 variables in homologue 23, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "25": contains 950 villages out of which only one village is a true outlier. As shown in figure 18, the number of flood events was the most important predictor and the outlier village had no flood events in the year 2013. The number of male migrant workers came second with the outlier village having none, however, in contrast to homologue "15", as the number of male migrant workers increases, Max EVI increases. The same was true for female migrant workers and the outlier village had few of them. Village revenue came fourth, with the outlier

village having a relatively low revenue. Figure 18 also shows that revenue sharing in managing wetland rice had an inversely proportional relationship with Max EVIs and the majority of rice households in the outlier village didn't use this type of land management. Pollution incidents also appeared as one of the important predictors of Max EVI and the outlier village didn't experience any. Drainage through sewage systems showed slightly lower Max EVI values than drainage through other systems. In season three, we can also see the majority of villages had a large proportion of households growing wetland rice and a very small proportion of households growing dryland rice. The outlier village had almost 90% of the households growing wetland rice and less than 20% of the households growing dryland rice in season three.

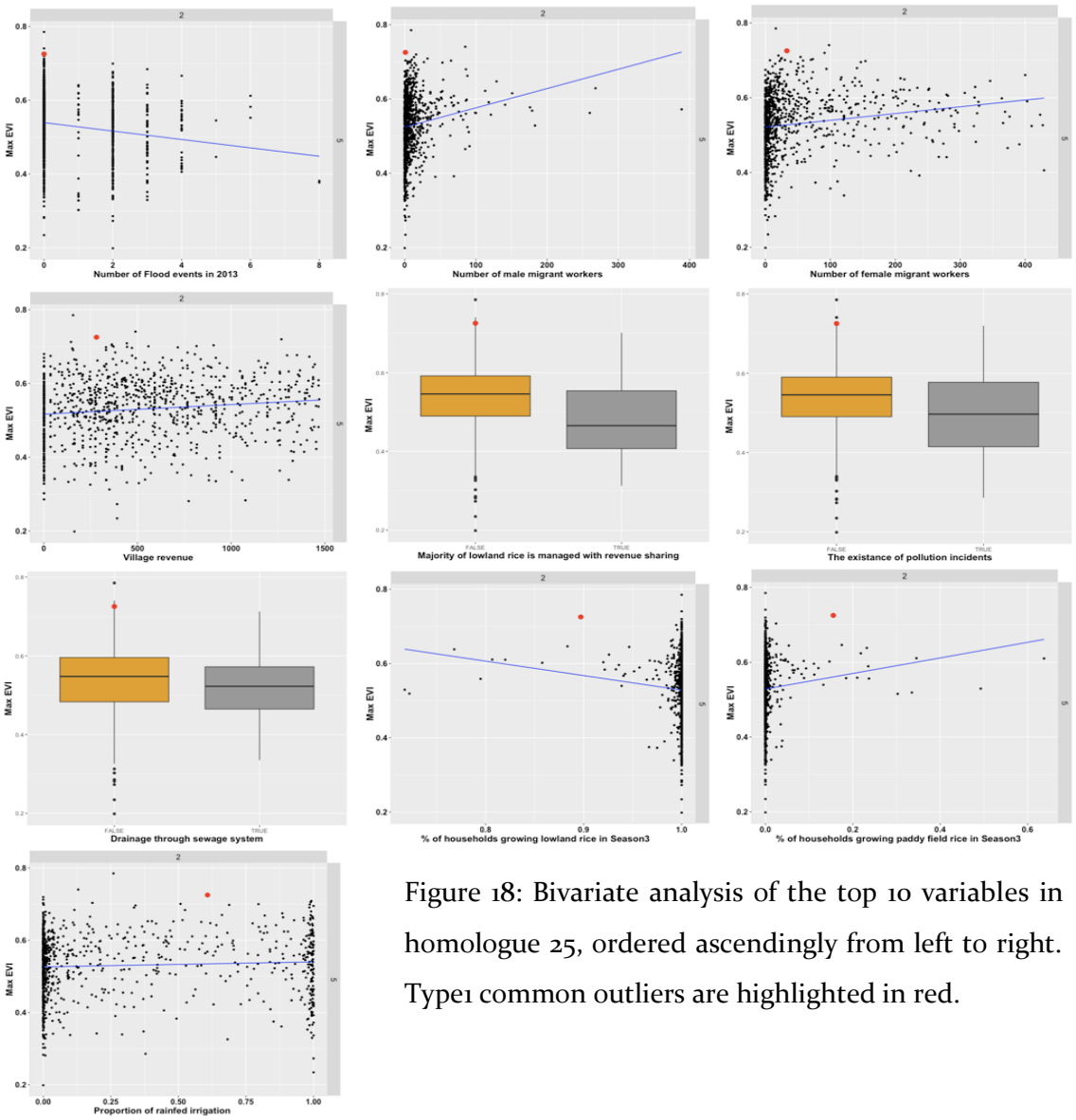

Figure 18: Bivariate analysis of the top 10 variables in homologue 25, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "32": contains 778 villages out of which only one village is a common true outlier. As shown in figure 19, the proportion of households growing dryland rice in season three is the most important predictor. However, the outlier village didn't have households growing dryland rice in season three despite having almost 70% of the households growing dryland rice in season two. On the other hand, almost all rice growing households in the outlier village grew wetland rice in season 3 and 90% of the rice households grew wetland rice in season two. The average age of the main farmer appeared again as an important predictor, with the

outlier village having an average age around 45. Opposite to previous homologues, the number of flood events here was directly proportional to Max EVI values. Figure 19 also shows that rain fed irrigation is one of the most important predictors of Max EVI, however none of the rice growing households in the outlier village used rain fed irrigation. Similar to the previous homologue, revenue sharing in managing wetland rice had an inversely proportional relationship with Max EVIs and the majority of rice households in the outlier village didn't use this type of land management.

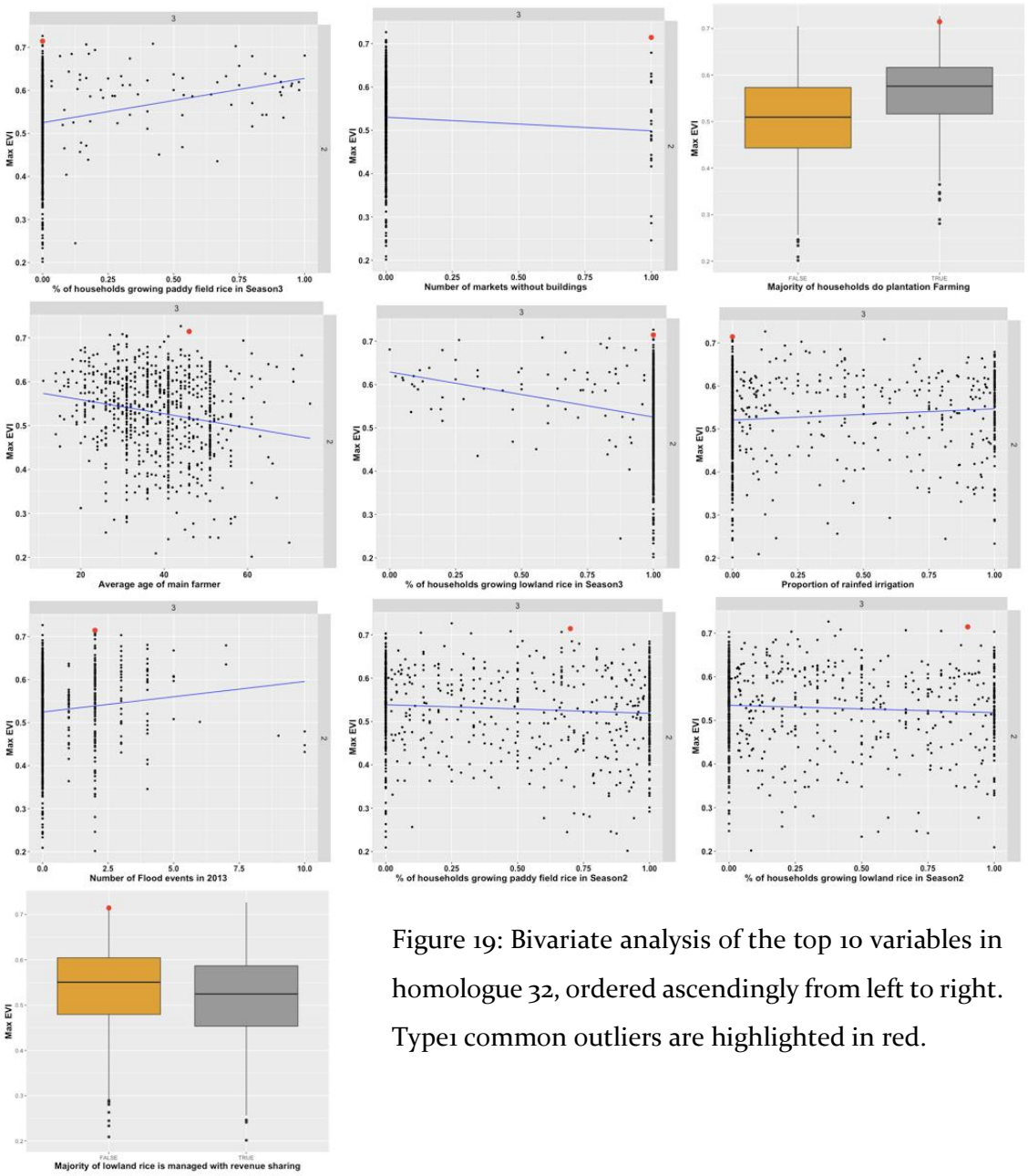

Figure 19: Bivariate analysis of the top 10 variables in homologue 32, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "33": contains 2265 villages out of which 13 villages are true outliers. As shown in figure 20, practicing plantation farming is again the most important predictor of Max EVI, however two outlier villages did not undertake plantation farming. Active saving and loan cooperatives appeared again with the majority of outliers having one or zero cooperatives. The number of families without electricity appeared also as one of the most important predictors. However, the majority of outliers had a very low number of families without electricity. In season three, the majority of villages - including outliers - had a very high proportion of rice households growing wetland rice and a very low proportion of rice households growing dryland rice. Having plantation farming as the main source of income and the main type of household business for the majority of households in the village, was identified among the most important predictors of Max EVI. Figure 20 also shows that outliers had varying proportions of households depending on rain fed irrigation in growing rice. The average age of the main farmer appeared again as an important predictor with outlier villages ranging from 30 to 50 years old. Finally, the number of landslides seemed to be positively correlated with Max EVI values and outliers had either one or zero landslides in the year 2013.

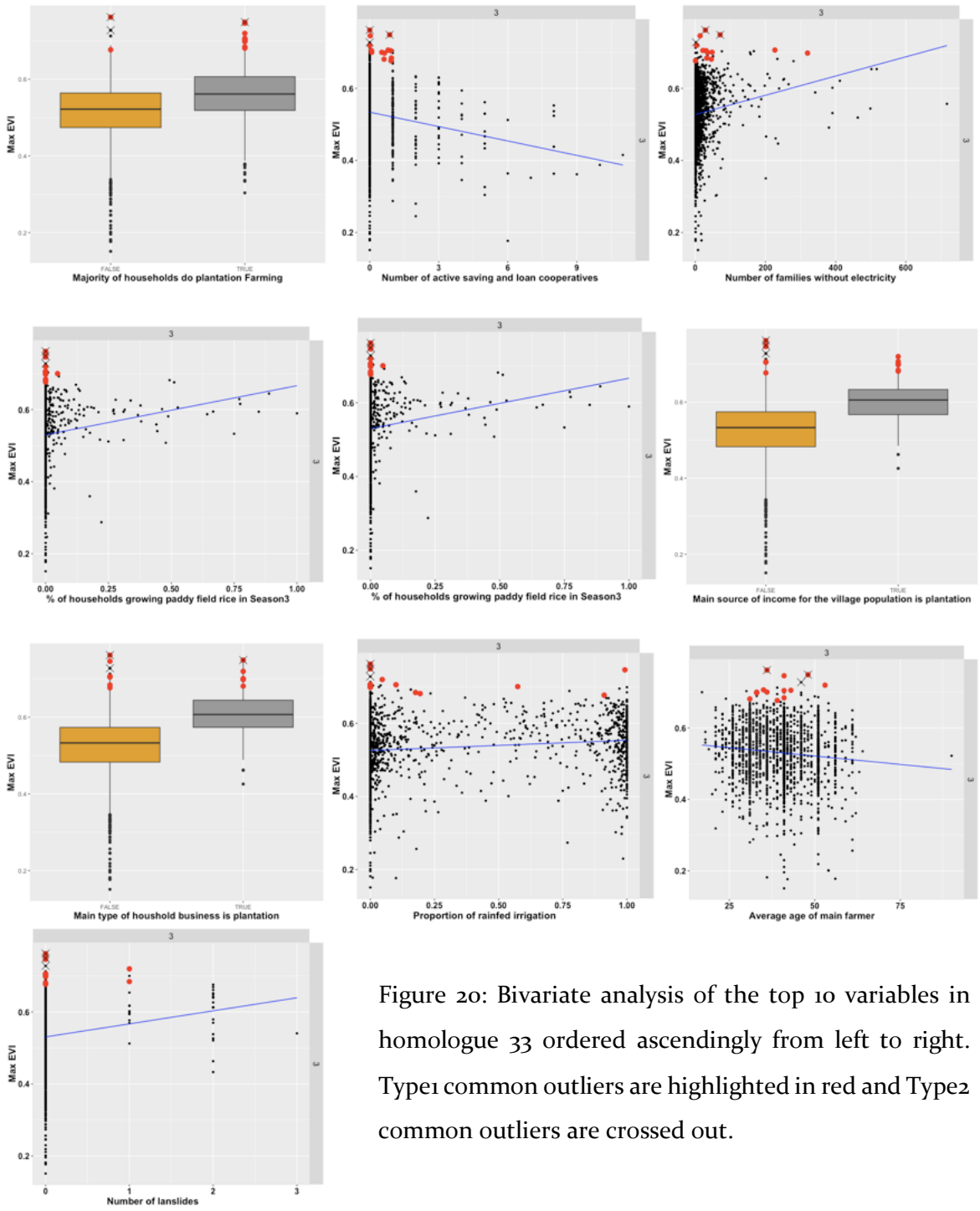

Figure 20: Bivariate analysis of the top 10 variables in homologue 33 ordered ascendingly from left to right. Type1 common outliers are highlighted in red and Type2 common outliers are crossed out.

HE "34": contains 1078 villages out of which three villages are common outliers. As shown in figure 21, doing plantation farming is also the most important predictor of Max EVI, however, one of the three outlier villages did not do plantation farming. Markets without buildings and active saving and loan cooperatives appeared again as an important predictor, however, all three outlier villages didn't have any of those markets and cooperatives. The proportion of rice households with rain fed irrigation in outlier villages is very low. Similarly, there were very few families without electricity in outlier villages. Average age of the main farmer in outlier villages ranged from 25 to 50 years old. In figure 21 burning fields to prepare the agricultural land appeared as an important predictor for the first time in this homologue and it is associated with better Max EVI values, two out of the three outlier villages did this practice. Other predictors that also appeared in this homologue and they might be related to its geographical location is having an area in the village that borders with the sea and the utilization of the sea for public transportation, the former was associated with lower Max EVI values and the latter was associated with higher Max EVI values.

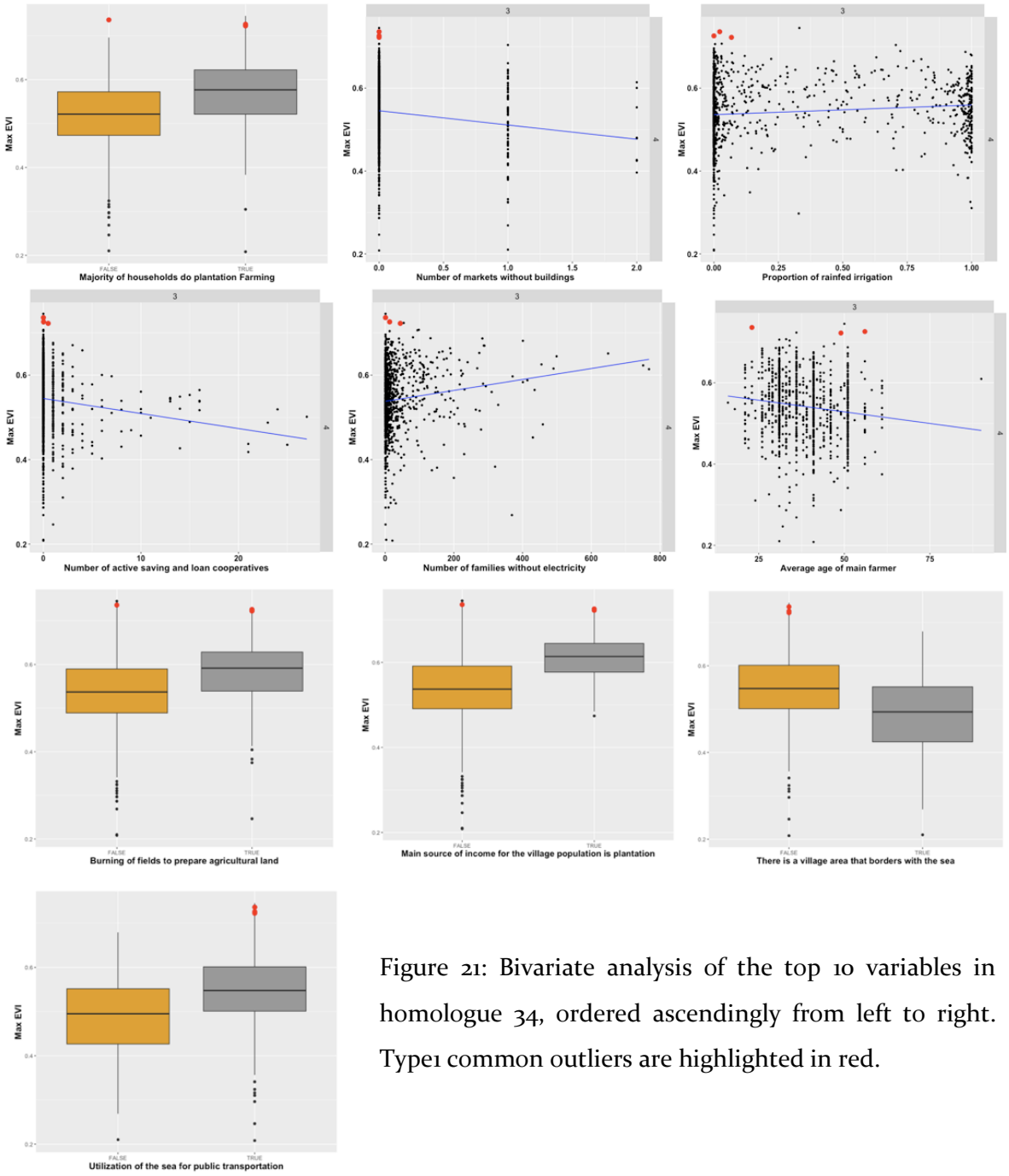

Figure 21: Bivariate analysis of the top 10 variables in homologue 34, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

HE "35": contains 520 villages, of which two villages are true outliers. As shown in figure 22, the proportion of rice households growing dryland rice in season two is the most important predictor of Max EVI with both outliers having 50% and 90% respectively. The use of LPG as the main cooking fuel also appeared as one of the top predictors and it was true for one of the outlier villages. Markets without buildings and families without electricity appeared again as important predictors, however, the two outlier villages had none. The average age of the main farmer for both the outlier villages ranged from 30 to 40. Figure 22 also shows that the number of female migrant workers was negatively correlated with Max EVIs, however the outlier villages had a very low number of those migrants. Doing plantation farming appeared again as a top predictor, however, the majority of rice households in outlier villages did not do plantation farming. In season three, the majority of villages - including outliers - had a very high proportion of rice households growing wetland rice and a very low proportion of rice households growing dryland rice. The number of flood events also appeared as an important predictor that is negatively correlated with Max EVI and outlier villages didn't have any flood events.

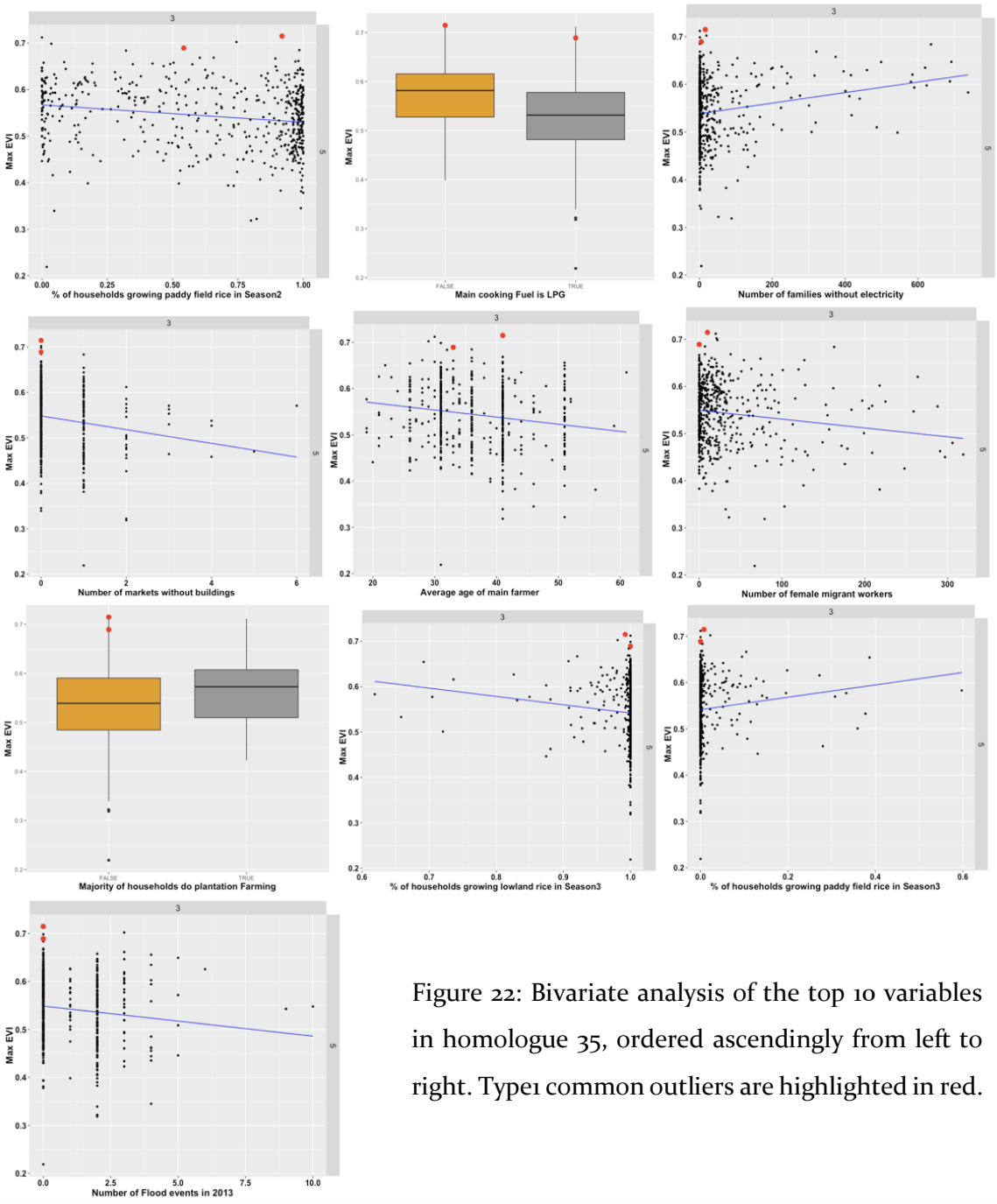

Figure 22: Bivariate analysis of the top 10 variables in homologue 35, ordered ascendingly from left to right. Type1 common outliers are highlighted in red.

Summary of Findings:

– In some of the HEs, it appears the absence of rice cultivation in the second season is associated with higher average Max EVI values. This is supported by evidence from a

study by Lantican et al. (1999) which suggests that farmers who grew non-rice crops before the rice season had better weed control, hence, improved rice productivity.

- We found predictors that were specific to certain homologues: for e.g. aquaculture household activities and horticulture farming in HE "14"; the existence of settlements in HE "23"; burning of the field in preparation of the agricultural land in HE "34"; and pollution incidents in HE "24". These results further add credence to the fact that production systems are complex, and highly varied, and the use of standardized administrative and EO data, with a PD approach, can help identify location specific constraints, and opportunities to improve agricultural performance.
- The results suggest that absent the use of rice type (i.e. dryland and wetland rice) as a control variable, there were no issues observed with identifying true PD rice producing areas across both types. Because, as shown in HE "13" of the bivariate analysis, out of the four outlier villages identified, two had the majority of households growing dryland rice and two had the majority of households growing wetland rice and still both were identified as outlier villages in the same homologue.
- The age of the farmer seemed to be negatively correlated with average Max EVIs. This supports the findings of Osanyinlusi & Adenegan (2016) who conducted a study that examined the factors affecting rice farmers' productivity of 160 randomly selected farmers in Nigeria . They offered evidence that farming experience was negatively significant to farmers' productivity. True outliers had an average age around the 30s.
- The bivariate analysis also shows that flooding events affect Max EVIs negatively. This is supported by the findings of Osanyinlusi & Adenegan (2016) which identified flooding as one of the constraints limiting rice production. Our true outliers had a small number of flooding events ranging from 0 to 3 floods in the year 2013.
- Type1 PDs didn't necessarily have the same values across the most important variables, on the other hand, type 2 PDs had clearly more conformity i.e. whenever they existed in a homologue, they had similar values across the various predictors of performance. This indicates that type 2 outlier detection doesn't only result in a smaller number of outliers, but also very similar outlier villages. This is evident in HE "14" & "33" of the bivariate analysis.

- Despite the positive relationship between the number of families without electricity and the average Max EVI values, the true outliers always appeared at the lower end with zero to few families without electricity.
- Other interesting findings from the bivariate analysis include the burning of rice fields to prepare the agricultural land and the existence of wells as bathing sources, which are both positively correlated with average Max EVI values. While we cannot explain the relationship between agricultural productivity and having wells as bathing sources, existing literature (Mandal et al. 2004) has shown that rice straw burning returns a considerable amount of plant nutrients to the soil in rice-based crop production systems.
- Despite the fact that "doing plantation farming" was listed as the top predictor of average Max EVI in a number of HEs, this doesn't necessarily mean that plantation farming is the main source of income for the village or the main type of business for the majority of households in this village. They could be villages that are growing other forms of plantation along with rice. For example in HE "33" we were able to identify 13 true outliers, from figure 9 we can see that 3 out of the 13 villages don't do plantation farming and 10 do plantation farming. Out of those 10 villages, only 4 had plantations as the main source of income and the main type of household business.

Google Time Scale Tool

We conducted a rapid check to investigate potential PDs and differences from true outliers using Google earth time scale tool, which enables us to view imagery obtained as near as possible to the selected cropping cycle (i.e. January to April 2013). All the true outliers were inspected in addition to 28 negative outliers. The presence and absence of rice, forestry/plantation and urban land cover was marked for each selected village, as well as basic notes describing the land cover. We did not quantify the land cover percentage but observed optical trends that became apparent. Inspecting each village using this method cannot conclusively determine whether a true outlier is a positive deviant. However, it can help determine if further investigation should take place for particular villages, to determine if certain land covers are potentially causing errors and where there are inaccuracies in the rice mask. Few of the trends we were able to identify among true outliers are presented below:

True outliers often had mixed land use including forestry and agricultural plantation around and within the rice area. The influence of mixed vegetation on the EVI value is currently unclear, however different vegetation covers will have different levels of 'greenness' (Heute et al. 2002; Megue et al. 2019). Figure 23 presents a village, which is a Type1 and Type2 outlier in HE "33". The white boundary is the rice mask and within the boundary there are rice fields. There is a small amount of urban land cover within the rice mask on the left. On the right, there appears to be mixed vegetation cover. This current analysis cannot determine if this village is a true PD but we suggest the village should be investigated further.

Figure 23: Mixed land use village (Tulleng village in East Nusa Tenggara, HE 33. The imagery shown was obtained on the 9th of September 2012)

Several Type1 outliers appeared to have monocultures of rice within the village boundary. Figure 24 presents two of such villages. These villages should be of interest for future analysis, as the influence of other land covers is potentially minimal. Weru is both Type1 and Type2 outlier, the rice boundary nearly covers the entire village, with other land covers being absent, within and around the rice mask.

Figure 24: Monoculture rice framing villages

(Weru, Banten and Kubangkampil, Pandeglang, Java, HE 15. The imagery was collected on the 10th of November 2014)

Many Type1 and Type2 outliers had accurate rice masks, with minimal urban and forestry land cover within the dedicated rice mask. Figure 25 presents two of such villages which had an accurate rice mask, with minimal mixed land-use within. There may be vegetation planted on the sides of the rice fields but it isn't detectable in this imagery.

Figure 25: Villages with accurate rice masks (Padang Subur (top), South Suliwesi, HE 25 and Bonne-Bonne Village (bottom), West Suliwesi, HE 33. The imagery was collected on the 8th of January 2013)

Four true outliers did not appear to have any rice within the rice mask. Figure 26 presents two of those villages who appeared to have no rice growing in the dedicated rice area and they had forest cover instead.

Figure 26: Villages with no Rice

(Soga, Soppeng Regency, South Sulawesi, HE 32 and Tada Selatan, Central Sulawesi, HE 33. Imagery was collected on the 4th of August 2010 and the 25th of May 2013, respectively)

On the other hand, a large number of negative outliers had low average Max EVI values because within the rice mask there was large amounts of urban land cover, with minimal rice production. Figure 27 shows two such villages.

Figure 27: Negative outliers

(Cibaduyut Kidul, West Java & Wumialo Gorontalo both in HE 33. Imagery was collected on the 9th of September 2012 and on the 8th of December 2014 respectively)

Based on the Google time scale tool, out of the 32 identified true outliers, rice was present in 29 villages and was absent in the remaining three villages. The remaining three villages which had plantation/forestation cover that may have contributed largely to the high Max EVI values, for the cropping season (Megue et al. 2019). This does not necessarily mean that those villages do not grow rice, but more accurate rice mapping data and a better proxy for productivity are needed before any conclusions can be drawn about rice productivity. The remaining 29 villages either had rice only or rice mixed with forestry, plantation or natural vegetation.

Earth Observation and Time Series Analysis

Hypothesis:

For this approach, we built and tested the following hypothesis:

"For a village to be an outlier, it is necessary for its agriculture production systems to be independent of climatic patterns, which means, despite fluctuations in climatic patterns, productivity (measured as the average max Enhanced Vegetation Index (EVI), wherein max EVI value for each pixel is averaged across all pixels belonging to a village, for the January to April 2013 season) of outlier village should remain consistent. Alternatively, we assume that agricultural productivity is tightly linked to climatic patterns among non-outlier villages, and their productivity is significantly associated with climatic patterns and seasonality. The basis for this assumption is that, outlier villages have adopted approaches and practices, and have established production systems, that delink climatic patterns with productivity, whereas the non-outlier approaches have not, at least for the target season, i.e. January to April 2013."

Previous studies have shown that Enhanced Vegetation Index (EVI), works well as a proxy for agriculture productivity, especially for annual crop systems in the tropics. It is also a well-

known fact that agriculture productivity is linked to climatic factors such as temperature and precipitation. Therefore, based on the stated hypothesis, and the relationship between EVI and agricultural productivity, and agricultural productivity and biophysical factors, the proposed validation step consists of the following logic:

(a) The triangular relationship can be leveraged to construct a per-pixel relationship between EVI, temperature and precipitation, and the same can be modelled across time (from 2001, until 2012) for both outlier and non-outlier villages.

(b) From 2012, for the same set of pixels, the model can be used to predict the EVI, given the temperature and precipitation values, since 2012.

(c) Based on the stated hypothesis, the assumption is that in 2013, for pixels belonging to outlier villages, the observed EVI values are significantly higher than the predicted ones, and this observation is reversed in the case of pixels of non-outlier villages (the rest of the sample in the HE), wherein the predicted EVI values are either equal or below the observed values.

(d) It is important to note that this particular logic has been derived to primarily quantify the differences between outliers and non-outliers for the target season (i.e. January to April 2013) only. The model during tuning and training process, will be able to learn any "violations" in the stated hypothesis, between the outliers and non-outlier villages before the target season, however since we do not test for such potential violations in any other season other than the target season, the proposed validation method cannot be used to infer differences between outlier and non-outlier before the target season.

Method:

Using the stated hypothesis, and the logic, this additional validation step consisted of the following workflow:

Based on the PLS regression results, we identified homologous environments (HE) that had the highest explanatory power for EVI. HE 21 and HE 22 were selected as both the environments explained ~ 75% of the observed variation in EVI. In addition, both the selected HE's revealed outliers based on only the multivariate approach (under both Type1 and Type2

outlier detection methods), and not based on the univariate approach. If differences between the observed and predicted EVI are identified using the proposed time series based approach, those differences can be attributed to the factors identified using the PLS approach.

Pixels belonging to outlier villages within these two HEs were grouped into one class (total sample size equalling to ~ 650 pixels of 1 square kilometre each), while pixels belonging to non-outlier villages in the same HEs were grouped into another class (total sample size equalling to ~ 4500 pixels of 1 square kilometre each). ** All pixels are standardized to 1 square kilometre, see explanation in step e.* The fact that outliers in HE 21 and 22, belonged to only multivariate based analysis (under both Type1 and Type2 outlier detection methods), suggests that outliers, and the rest of the villages in HE 21 and 22 (non-outliers), are two independent, distinct samples, with different production practices and constraints, ~ 75% of which can be attributed to factors identified to be important differentiators between the two samples using the PLS approach. Therefore, comparisons were made within the pixels of outliers, and non-outliers, across time, rather than between outliers and non-outliers across time.

Hence, the following steps were conducted separately for both classes, and comparisons were performed across time within each class.

a) For each pixel, monthly data since January 2001, until December 2016, for temperature, precipitation and EVI was obtained. The temperature data was obtained from MODIS's Land Surface Temperature and Emissivity sensor (MOD11C3), that provides global monthly day time land temperature, at ~ 5 square kilometre resolution. Similarly, for the same time period (i.e. between 2001 January and December 2016), monthly per pixel EVI values were obtained from MODIS terra sensor (MODIS VI), at 1 square kilometre spatial resolution. Monthly average precipitation data for the same set of pixels and for the same time period, were obtained from the CHIRPS (Climate Hazards Group Infrared precipitation with Station Data) database at ~ 5 square kilometre resolution.
b) In order to bring temperature and precipitation data to the same spatial resolution as the EVI data, temperature and precipitation raster stacks (stacked across time) were

resampled using the resample algorithm of Raster Package in R statistical environment.

c) Spatial and temporal gaps across the three datasets were assessed using the Amelia Package in R, and a simple moving average approach from ImputeTS Package in R statistical environment was used to fill data gaps across time.

d) For each class (i.e. outliers and non-outliers) separately, monthly data from January 2001, until December 2012, was used for model building and validation, wherein EVI was used as the response variable, while precipitation and temperature were used as predictors. Separately for each class, monthly data for temperature and precipitation since January 2013, was provided to the model, in order to obtain predicted EVI values.

e) For the modelling approach, we relied on the grid search capabilities, and a deep learning model (a feed-forward multilayer perceptron) provided by the in R statistical environment.

f) For both the models for each class, hyper-parameter tuning was conducted using a random-discrete based grid search approach, with rectifier and Maxout, with and without dropout as activation functions. The grid search approach also included three different combinations of hidden nodes, and a range of values for lasso and ridge regularization (to prevent overfitting). Lastly, the grid search was restricted to testing for a total of 20 models, with 5 folds and 100 epochs.

g) The best performing deep learning model was selected using RMSE and MAE validation metrics.

h) The best model for both the classes, was used to perform predictions. For the predictions, the best performing model was provided with monthly temperature and precipitation data from 2013 January onwards, until December 2016.

i) Package CAST in R statistical computing environment was used to build spatiotemporal cross validation folds, to obtain training and validation datasets for model building purposes. Two folds across space and time were derived from the training data separately for the two classes. Fold one was used to build the model, while the other fold was used as the validation dataset.

j) For the modelling approach, temperature, precipitation data was normalized to scale between 0 and 1, using the normalize function. In addition, the EVI data was also

rescaled to 0 and 1. Data rescaling was done using the Normalize function in the BBMISC Package in the R statistical environment.

k) Scatter plots for comparing the predicted EVI values and scaled and observed EVI values from the validation dataset were constructed using the ggplot2 plotting Package in R statistical environment.

l) A new variable called difference (diff) was constructed, which quantifies the difference between Observed (and scaled), and predicted EVI values, for each pixel, separately for both outlier and non-outlier villages. Histograms for the difference values were constructed using base R functions, and density plots for the same were constructed using the ggplot2 Package in R statistical environment.

m) For true outlier and non-outlier villages separately, total number of positive and negative difference values were counted. Positive difference values reflect that the observed is higher than the predicted, while the negative difference values correspond to pixels, wherein the predicted values are higher than the observed.

n) Count data for positive and negative values was subjected to chi-square analysis, in R statistical environment. The null hypothesis in this case referred to an equal number of positive and negative values. Chi square test was done separately for true outlier and non-outlier villages.

o) Since the classification of outlier and non-outlier villages was performed with village as the observational unit, we performed additional analysis, to reflect pixel level differences in observed and predicted EVI values, to the observational unit.

p) For the village level analysis, EVI raster layer, in which pixel values were masked using spatial polygon shapefiles of either outlier or non-outlier villages, wherein each polygon belongs to a village, was used. Each pixel was converted into a polygon using the Raster Package in the R statistical environment. Following this, a cell number was assigned to each polygon.

q) The cell number in the above step, corresponds to the cell number of pixels for which difference values were calculated.

r) Original values for cells in the EVI raster layer were replaced with the corresponding differenced values of the corresponding cells.

s) The difference values in the EVI raster layer were again extracted, this time into the corresponding village, using the spatial polygon shapefiles of either outlier or non-outlier villages.

t) For each village, the total number of positive and negative difference values were counted, and the counts were subtracted (referred to as diff2 in the figures)

Results and Discussion:

The model building and validation employed in this approach, relies on the established knowledge regarding the relationships between (a) biophysical covariates, i.e. Temperature and Precipitation, and EVI, (b) relationship between EVI and agricultural productivity, and (c) biophysical covariates and agricultural productivity. Therefore, although two different models, one for outlier village pixels, and the other for non-outlier village pixels were developed, the strategy was to select a model, which best reflects this existing knowledge of relationship. Therefore, the model building process for both the classes involved the same hyperparameter ranges and search strategy. Table 24 presents model parameters for outliers and non-outliers, which best describes the relationship between biophysical covariates and EVI. Interestingly model validation metrics, for both with the validation dataset, and the prediction dataset, showed that the model for pixels belonging to outlier villages had a better fit than the pixels of non-outlier villages (See RMSE and MSE values in Table 24). The number of pixels in the non-outlier villages were significantly much more than those in the true outlier villages. Therefore, pixels in non-outlier villages could encompass larger variation in biophysical covariates and EVI, compared to those in the outlier villages, hence rendering it difficult to find a model that best describes the relationship in comparison to pixels in outlier villages, which is reflected in the model validation metrics. The focus of this modelling strategy is to perform predictions on data from the year 2013, and not to describe the relationship between biophysical covariates and EVI, however it is interesting to note that EVI is significantly influenced by precipitation in pixels among non-outlier villages, while EVI is influenced by temperature among pixels in outlier villages.

Best Model Specification (obtained after grid search)	**Non-Outlier villages**	**Outlier villages**
Activation	Rectifier	Max out with dropout
Hidden neurons	20, 15	20, 15
l1 (lasso optimization)	1.0E-5	0.001
l2 (ridge optimization)	1.0E-5	0.001
Mean squared error (MSE) on validation data (Fold 2)	0.012	0.0005
Root mean squared error (RMSE) on training on validation data (Fold 2)	0.111	0.02
Top important variable	Precipitation	Temperature
Mean squared error (MSE) on prediction data	0.015	0.002
Root mean squared error (RMSE) on prediction data	0.126	0.05

Table 24: Selected parameters and validation metrics for the best model obtained after employing the same search strategies, across identical ranges of hyper-parameter values for pixels belonging to both the classes, i.e. outlier and non-outlier villages

Prediction results from the selected models for both classes, also reflect the model validation metrics presented in Table 24, wherein the slope of the relationship between observed and predicted scaled EVI values for pixels belonging to outlier villages, differed significantly to those from the non-outlier villages. Relatively steeper slope, in the case of pixels belonging to outlier villages, suggests that in general, the observed EVI values for the period January to April 2013, are significantly higher than the predicted values, unlike those among pixels of non-outlier villages (Figure 28). This further suggests that production systems belonging to

pixels in outlier villages are performing better than expected (i.e. predicted, and are not entirely dependent on climatic conditions).

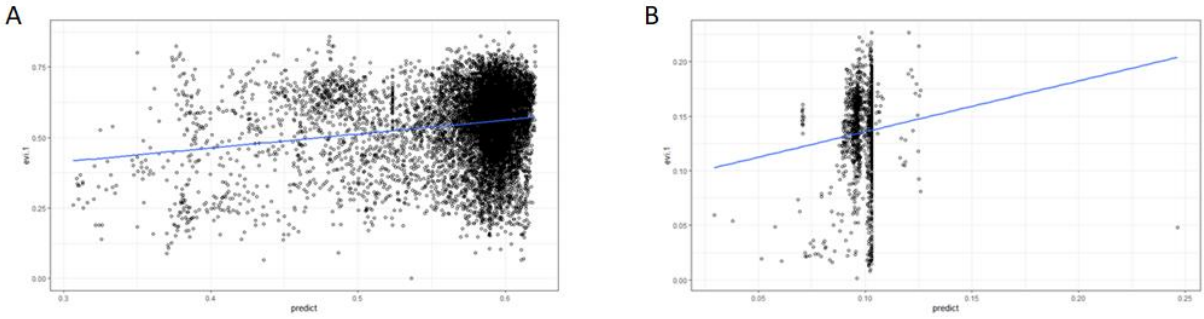

Figure 28: Scatter plot presenting prediction results from the best model, using monthly values of biophysical covariates starting from January to April 2013 (i.e. the season for which data from the Agriculture census was used to identify true and non-outlier villages). X axis represents the predicted EVI values (scaled), while the Y-axis represents the observed EVI values (also scaled) for each pixel, between January and April 2013. The blue line in each scatter plot represents a linear model fit. Subset A represents the predictions for pixels belonging to the non-outlier villages, while B represents the same for pixels belonging to outlier villages.

The observation that pixels in outlier villages perform better than expected, is further evidenced by the fact, that the distribution of the difference values for pixels (i.e. per pixel difference between scaled values of observed and predicted EVI), is skewed to the right (Figure 29 C and D), indicating presence of more number of positive difference values, in comparison to pixels in non-outlier villages, wherein no skew was observed (Figure 29 A and B), indicating relatively equal numbers of positive and negative difference values. More number of positive than negative difference values, as observed in the case of pixels of outlier villages, suggests that many pixels performed better than expected, during the January to April 2013 season.

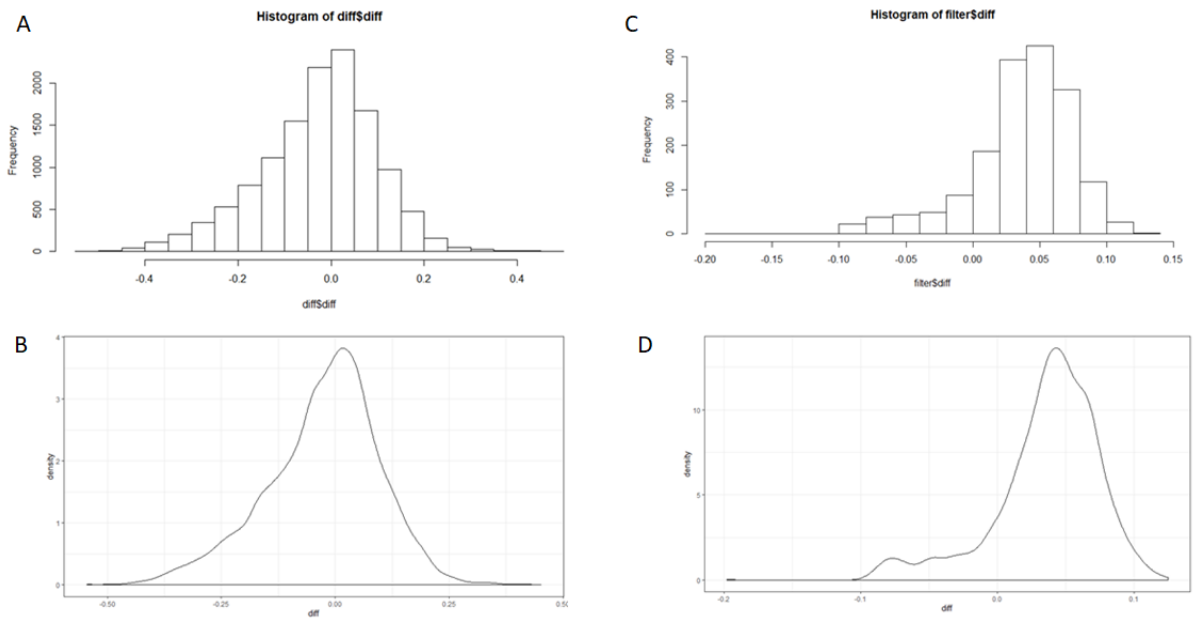

Figure 29: Distribution of the difference values for pixels

(Histograms and density distribution plots representing the distribution of differenced values (difference between observed and predicted scaled EVI values) across all pixels belonging to non-outlier villages in subsets A and B, and outlier villages in subsets C and D)

In order to test if the difference in the number of positive and negative difference values, for either pixels of and non-outlier villages, is not a chance observation, we performed a chi-square test, with the null hypothesis of equal proportion of positive and negative difference values.

Chi square test for pixels in both outlier and non-outlier villages showed that the difference between positive and negative difference values is significant, and the observation is not by chance (Table 25 and Table 26; chi square statistic for non-outlier villages = 48.099, chi square statistic for outlier villages = 512.4256, p-value < .01). However, the number of positive difference values are significantly higher than negative values in outlier pixels, in contrast to that of the pixels in non-outlier pixels, wherein the number of negative difference values are significantly higher than positive values, showing that pixels of outlier villages did indeed perform better than expected during the January 2013 to April 2013 cropping season.

Non-Outlier pixels	neg	pos	
Null Hypothesis	6324	6324	
Actual	6875	5773	
The chi-square statistic is 48.099			
The p-value is < 0.00001. Significant at p < .01			

Table 25: Chi-square test analysis for non-outlier villages

Table 25 presents Chi-square tests to assess if the difference between the total number of negative (neg) and positive (pos) difference values among pixels of non-outlier villages are significantly different, and are not observed by mere chance. Actual in the above table refers to the observed number of pixels with positive and negative difference values, while Hypothesis refers to the null hypothesis of equal number of positive and negative difference value pixels.

Outlier pixels	neg	pos	
Null Hypothesis	860	859	
Actual	240	1477	
The chi-square statistic is 512.4256			
The p-value is < 0.00001. Significant at p < .01			

Table 26: Chi-square test analysis for outlier villages

Table 26 presents Chi-square test analysis to assess if the difference between the total number of negative (neg) and positive (pos) difference values among pixels of outlier villages are significantly different, and are not observed by mere chance. Actual in the above table refers to the observed number of pixels with positive and negative difference values, while

Hypothesis refers to the null hypothesis of equal number of positive and negative difference value pixels.

Since the identification of outliers and non-outliers was performed at the lowest administrative unit (i.e. village) in the agriculture census, the results from the time series validation were rolled up from the pixel level to the village level. Distribution of the difference2 values (i.e. difference between the number of positive and negative difference values) across all villages, belonging to either outlier or non-outlier class revealed that a major proportion of villages in the case of outlier (18 out of 19; Figure 30B) villages obtained positive difference2 values. This is in contrast to non-outlier villages, wherein relatively lower proportion of villages (251 out of 580; Figure 30A) obtained positive difference2 values, further indicating that, aggregation of the pixel results to village level, also shows that indeed villages belonging to true outlier class performed better than expected, during the cropping season between January 2013 to April 2013.

These results also reveal that the per pixel and village level differences observed between outliers and non-outliers, can be attributed to the factors, identified from the PLS regression, responsible for differentiating the two village types.

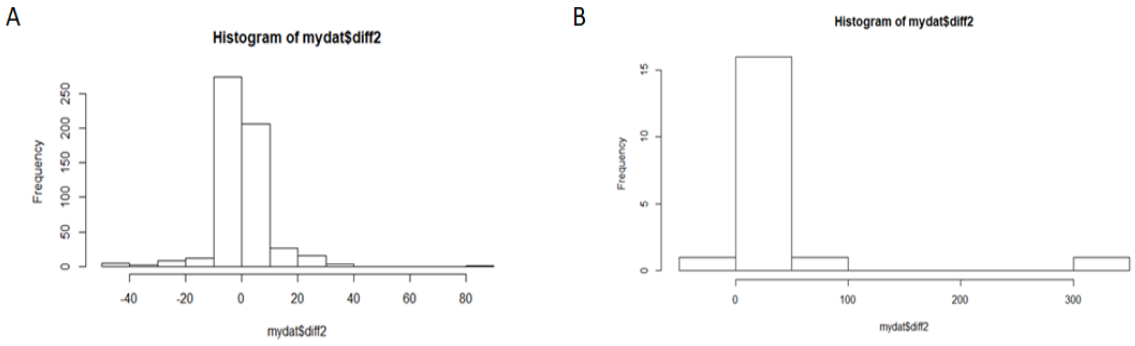

Figure 30: Histogram presenting the distribution of positive and negative difference2 values (i.e. the difference between the number of positive and negative difference values) aggregated at the village level for pixels belonging to (a) Non-Outlier villages (b) True Outlier villages

Histogram analysis of difference2 values showed that Sungai Lumpur village (Geocode: 1602031004), in Sumatra Selatan province, had the highest difference2 value (= 349) among true outlier villages, while Barabai Utara village (Geocode: 6307050006) in Kalimantan Selatan province, was the only village with a negative difference2 value (= -1) among true outlier villages. Interestingly, histogram analysis also revealed several villages with positive difference2 values. Tirto Sari village (Geocode: 1607060026) in Sumatra Selatan, had the highest positive difference2 value (= 88) among non-outlier villages. In contrast Bintaran village (Geocode: 1607091004) in Sumatra Selatan province, had the highest negative difference2 value (= -49), suggesting that, among all the non-outlier villages, this village performed the worst, during the January 2013 to April 2013 cropping season. These results also reveal an interesting observation with respect to province Sumatra Selatan, as both the best and worst performing villages, were identified in the same province.

In addition to validating the identification of true outlier and non-outlier villages, this step can also be used to further narrow down the number of villages for additional ground truthing, to identify true positive deviants. It is important to note that this method needs further development to test which methods (multivariate or univariate - or Type1 and Type2 outlier detection approaches), yield true PDs.

4.5 Challenges and Limitations

- **Rice mask errors:** To minimize the influence of other land covers on the extracted EVI values, we used the intersection of the rice mask from the 2014 Indonesian Land Use shapefile and the village boundaries. However, the 2014 Land Use shapefile (provided by the Ministry of Forestry) only indicates the rice area as "Sawah" (Indonesian for "rice field") and does not differentiate between "wetland rice" and "dryland rice." It is unknown what rice variety is grown in the mask and it cannot be directly paired to the census and PODES datasets. Additionally, as discussed in the section earlier on validation using Google Earth, the rice mask boundaries are not accurate, likely because it was created with a visual classification with medium resolution data (Setiawan et al. 2013). When viewing the land cover of villages, rice can be found outside the dedicated rice mask and alternative land covers, such as urban land, industry, and other agricultural systems are also often found within.

- **Dataset integration errors:** We identified a number of potential sources of error when combining the different datasets. The first potential source of error identified, was the false geometry in the village administrative boundary shapefile. To combine EO data with the census data, there is a crucial step of extracting the raster values using the administrative boundary shapefile. It was assumed that correcting the geometries of the shapefiles would reduce the errors and decrease the reduction in the sample size. Whilst correcting geometry improved the results of the extraction, there was still a reduction in the sample size. This could be due to two reasons. First, the extraction method is failing because the spatial resolution of the rasters in comparison to the administrative boundaries are too large. Second, the spatial data have inherent errors coming from the coordinate reference systems, data frames, and extent. Whilst we corrected and pre-processed the data, there were still errors when combining the data. In the future, different extraction methods such as binary, centroid or interpolation methods should be trialled.
- **EO data errors:** When extracting the CHIRPS data into the 78,811 villages there were a total of 1,328 villages with NA values. When displayed spatially, the NAs are mostly present along the coast, on small islands amongst the archipelago and near inland lakes. And when we attempted to extract the temperature raster onto the 78,811 villages, there were 620 villages with NA values. When displayed spatially we found an overlap with the errors previously displayed with the CHIRPS data. There was an additional source of error when we extracted the average EVI Max values for rice growing areas within villages. Across the administrative boundary of a village, there are numerous rice growing areas that return null values and negative values. Of the 36,787 villages, there were 1,194 villages with no value or a negative value from the raster. Figure 31 presents a map showing the spatial location of those errors, with insets displaying randomly selected locations at a higher resolution.

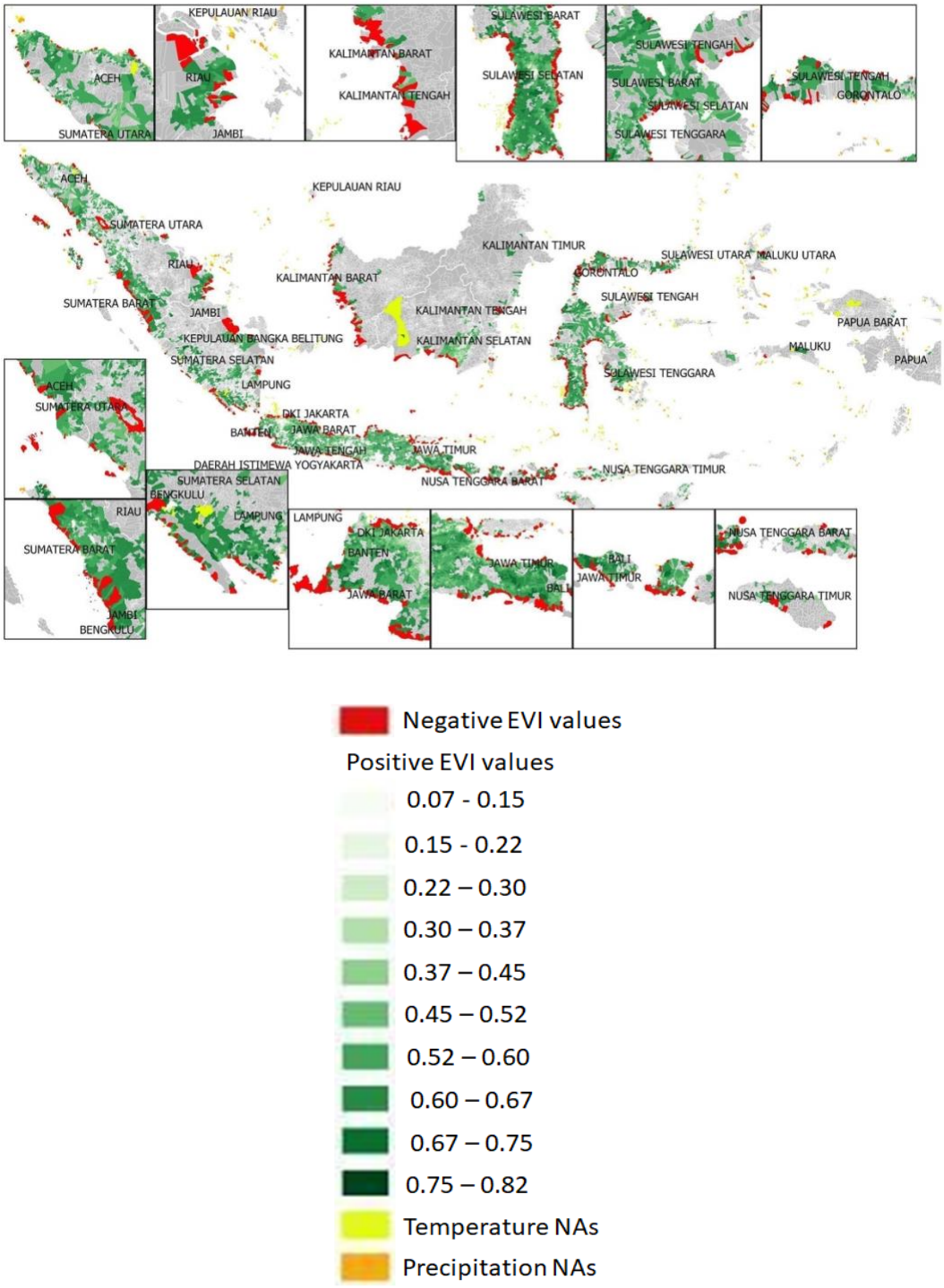

Figure 31: The location of missing EVI, precipitation and temperature data across Indonesia, after extracting the raster values with the village administrative boundaries

- **Complex land cover errors:** Where the rice land use area is relatively small and surrounded by forest or plantations, there could be a potential source of error resulting from the MODIS imagery, which has a moderate spatial resolution of 250 metres per pixel (Setiawan et al. 2013). There is a possibility that the EVI value extracted from the rice area is influenced by the spectral signature of the surrounding land use (Setiawan et al. 2013). Given the complexity of the landscape, it is possible that other high performing villages have low EVI values due to the spatial proximity of low reflecting land covers, i.e. urban settlements, roads and barren areas. For example, figure 32 presents two villages that are potentially subject to such errors. In the village to the left, the rice mask within this village is mostly rice, with some other vegetation within. However, the rice growing area is surrounded by forestry and native vegetation. It is unknown if the spectral signature of the surrounding land cover influences the EVI value within the rice mask. Whereas in the village to the left, there is a large amount of urban and industrial land cover within and surrounding the rice mask, potentially reducing the EVI value.

Figure 32: Examples of complex land cover errors (On the left is West Teupah Village, Aceh (HE 13). The imagery shown was obtained the 5th of March 2014. The dedicated rice growing area in this village is mostly rice, with some other vegetation within. On the right, Sukaurip, Indramaya Regency (HE 14). The imagery was collected the 17th of August 2013)

4.6 Recommendations for Future Work

- **Performance measure:** The performance measure used in this study, i.e. average Max EVI, was calculated by extracting the Maximum EVI value for each pixel, within a village boundary, and averaging the Maximum value, across all pixels within this boundary. However, several nuances of rice crop phenology across a season are missed in this performance measure. For instance, using temporal vegetation indices and biophysical covariates data, for pixels within the rice mask, it is possible to predict, and construct a performance measure that captures multiple characteristics of rice crop phenology for the target season, such as initiation of greening (beginning of season), senescence (end of season), length of season etc, all of which can be captured per pixel, and then be aggregated to the rice mask for each village.
- **Studying negative outliers:** With PLJ's focus on researching and developing innovative tools and approaches to promote inclusive growth that leaves no one behind, future investigations should include the negative outliers as well. Comparing the practices and enablers of positive deviants with those of negative deviants is crucial in designing effective interventions and for identifying the key elements responsible for higher productivity.
- **Rice area mapping:** Instead of using rice crop masks that are inaccurately separating areas growing rice from other land covers and land use, alternative methods could be used for rice area mapping. One of the methods is using MODIS multi-temporal satellite imagery to map rice areas. A study by Lee et al. (2012) produced a map of dryland distribution in Indonesia. The algorithm they developed uses time series of various vegetation indices (i.e. EVI and NDVI) to identify the initial period of dryland flooding and transplanting based on the sensitivity of the land surface water index (LSWI).
- **Control variables:** In this analysis, we only use data from biophysical variables that were open source and readily available. There is a need to control for other drivers of agricultural productivity that farmers have no control over (such as soil) by engaging with agricultural experts, that could suggest such variables.
- **Expert driven validation:** Administrative/government agencies, who in this case are also data collectors, will be the end users of the results obtained from the proposed PD approach. Therefore, it is necessary to validate these results, through focus group

meetings and consultations to understand: 1) If indeed the variables identified with the PLS/Bivariate analysis, differ between true outliers from the remaining villages in the same HE; 2) if these are also determinants of higher average Max EVI values among true outliers and 3) in comparison to the rest of the outliers, and in general if agricultural performance among true outliers is better than other villages in the same HE.

- **Complex land covers:** Land cover across Indonesia, including agriculture is complex, as 80% of 250 m pixels are not homogenous (Setiawan et al. 2011) and we suspect that there are potential influences from the reflectance of multiple land covers within a MODIS pixel (Setiawan et al. 2013). Obtaining higher spatial resolution data land use data and (Setiawan et al. 2011; Setiawan et al. 2014). In the future, it may be worthwhile to explore Landsat imagery with a 30 metres spatial resolution and other imagery to obtain more accurate agricultural mapping (Mengue et al. 2019).
- **Google earth validation:** The open sourced imagery available in the Google Earth Time Scale Tool does not have a consistent time stamp across Indonesia. Some areas have more available imagery and some areas of interest did not have imagery available during the target time period. We consistently tried to inspect imagery with a collection data as close as possible to our target time period to the chance of land cover change but due to the time difference. By inspecting imagery as close as possible, we assume that the land cover, especially the rice and forest cover did not change between the date of collection and our target date.

4.7 Conclusion

The positive deviance (PD) approach for development programming relies on identifying and scaling the strategies of positive deviants (PDs) i.e. individuals or communities who use uncommon practices and behaviours that enable them to achieve better outcomes than their peers. Disseminating and analysing the behaviours and other factors underpinning PDs are demonstrably effective in delivering development results. However, conventional PD approaches are time and labour-intensive, and often are not scalable across communities. This is because they rely mainly on primary data collection for the identification of PDs with costs proportional to the sample size. Hence, the samples are usually small, which can make it hard to identify PDs statistically and practically, given their relative rarity. Innovations in

digital technologies and platforms that record, mediate or observe individual and community behaviours, led to the proliferation of digital datasets "big data" (e.g. online data, mobile data and earth observation data) that could enable us, in specific domains, to identify and understand PDs in new and/or better ways (Albanna & Heeks 2019). In agricultural development programming, earth observation (EO) data have the potential to provide deeper insights on behaviours of rural communities at temporal and spatial scales not previously possible using conventional methods.

In this study we presented a stepwise approach for the identification and validation of PD rice villages in Indonesia i.e. villages with significantly higher agricultural productivity in comparison to neighbouring villages with similar socio-economic and environmental conditions. It is a step towards building evidence for the use of big data to facilitate PD related development programming in agriculture. We were able to demonstrate that big data sources (such as EO data), can be combined with administrative data, in order to spatially locate and identify PD communities and some of their underlying practices such as straw burning, demographic variables such as average age, and contextual variables such as type of irrigation; which is a precursor to mainstreaming PD in development programming.

The presented analysis shows that the administrative data was able to explain variance in agricultural performance - captured through the EO derived measure - ranging from 21% to 75% across the 15 pre-determined homologues. This suggests that there are factors affecting performance that are not fully captured using the administrative data and some of those factors could be identified through extensive ground surveys and ethnographic methods. Collecting such data for large samples is difficult. However, in this study we provide a systematic way to identify information rich small samples - characterized in true outliers or PDs - that could be targeted for ground data.

We specifically focus on the use of administrative data, as we see national governments, as primary stakeholders. Through this study, we provide evidence that administrative data, when combined with open source Earth Observation data, can be reused in multiple ways to facilitate targeted development planning.

Although the advantage of big data is in its ability to provide more data, the trade-off however is that more data, also could incorporate more noise. Therefore, in this study, we focused on developing a statistically rigorous approach, with multiple validation steps, in our bid to separate noise from true insights.

The developed methodology can be used to draw other insights from the combination of open source EO with administrative data, that are not directly connected to PD. For instance, we identified 4 homologous environments (HEs) within Aceh province, whereas 3 HEs where in Java, yet the sampling strategy for the two provinces is the same in the Agriculture census, thereby leading to under-representation of diverse conditions and complexities in Aceh, within the census.

This proof-of-concept analysis contributes to the evidence that big data sources and analytics, administrative data, and open source EO data, has the potential to facilitate mainstreaming of PD into development programming, further empowering national and local governments, with methods that can enable targeted bottom-up solution development. However, the roll out of this method would require the use of recent administrative data along with EO data in order to move to the next stage of PD inquiry i.e. ground surveys and ethnographic methods targeting the true outliers or PDs to understand their underlying behaviours.

References

Albanna, B. & Heeks, R., (2019) Positive deviance, big data, and development: A systematic literature review, *The Electronic Journal of Information Systems in Developing Countries*, *85*(1), e12063.

Bolton, D. K. & Friedl, M. A. (2013) Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics, *Agricultural and Forest Meteorology*, 173, 74–84.

Cinner, J.E., Huchery, C., MacNeil, M.A., Graham, N.A., McClanahan, T.R., Maina, J., Maire, E., Kittinger, J.N., Hicks, C.C., Mora, C. & Allison, E.H., (2016) Bright spots among the world's coral reefs, *Nature*, 535(7612), p.416

de Vries, F.P. ed., (2005) Bright spots demonstrate community successes in African agriculture (Vol. 102). IWMI.

Hartini, T.N.S., Padmawati, R.S., Lindholm, L., Surjono, A. & Winkvist, A., (2005) The importance of eating rice: changing food habits among pregnant Indonesian women during the economic crisis, *Social Science & Medicine*, *61*(1), 199-210.

Heute, A., Didan, T., Miura, T., Rodriguez, E.P., Gao, X., & Ferreira, L.G., (2002) Overview of the radiometric and biophysical performance of the MODIS vegetation indices, *Remote Sensing of the Environment, 83,* 195 - 213.

Johnson, D.M., (2014) An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States, *Remote Sensing of Environment*, *141*, pp.116-128.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E. & Nizam, A., (1988) *Applied regression analysis and other multivariable methods* (Vol. 601). Belmont, CA: Duxbury Press.

Lee, N., Monica, A. and Daratista, I., (2012) Mapping Indonesian dryland using multiple-temporal satellite imagery, *African Journal of Agricultural Research* , *7*(28), 4038-4044.

Lantican, M.A., Lampayan, R.M., Bhuiyan, S.I. & Yadav, M.K., (1999) Determinants of improving productivity of dry-seeded rice in rainfed wetlands, *Experimental Agriculture*, 35(2), 127-140.

Maitra, S. & Yan, J., (2008) Principle component analysis and partial least squares: Two dimension reduction techniques for regression, *Applying Multivariate Statistical Models*, *79*, 79-90.

Mandal, K.G., Misra, A.K., Hati, K.M., Bandyopadhyay, K.K., Ghosh, P.K. & Mohanty, M., (2004) Rice residue-management options and effects on soil properties and crop productivity, *Journal of Food Agriculture and Environment*, 2, 224-231.

Mengue, V.P., Fontana, D.C., da Silva, T.S., Zanotta, D. & Scotta, F.C., (2019) Methodology for classification of land use and vegetation cover using MODIS-EVI data, *Revista Brasileria de Engenharia Agricola e Ambeintal, 23*(11), 812-818.

Mkhabela, M.S., Bullock, P., Raj, S., Wang, S. & Yang, Y., (2011) Crop yield forecasting on the Canadian Prairies using MODIS NDVI data, *Agricultural and Forest Meteorology*, *151*(3), pp.385-393.

Osanyinlusi, O.I. & Adenegan, K.O., (2016) The determinants of rice farmers' productivity in Ekiti State, Nigeria, *Greener Journal of Agricultural Sciences,* 6(2), 49-58.

Pant, L.P. & Hambly Odame, H., (2009) The promise of positive deviants: bridging divides between scientific research and local practices in smallholder agriculture, *Knowledge Management for Development Journal*, 5(2), pp.160-172.

Qiu, B., Zeng. C., Tang, Z., & Chen, C. (2013) Characterizing spatiotemporal non-stationarity in vegetation dynamics in China using MODIS EVI dataset, *Environmental Monitoring and Assessment*, *185*(11), 9019-9035.

Setiawan, Y., Yoshino, K. & Philpot, W.D. (2011) Characterizing temporal vegetation dynamics of land use in regional scale of Java Island, Indonesia, *Journal of Land Use Science,* 8(1), 1- 30.

Setiawan., Y., Yoshino., K. & Prasetyo., L.B., (2013) Characterizing the dynamics change of vegetation cover on tropical forestlands using 250m multi-temporal MODIS EVI, *International Journal of Applied Earth Observation and Geoinformation, 26,* 132-144.

Son, N.T., Chen, C.F., Chen, C.R., Minh, V.Q. & Trung, N.H., (2014) A comparative analysis of multitemporal MODIS EVI and NDVI data for large-scale rice yield estimation, *Agricultural and Forest Meteorology*, *197*, 52-64.

Sternin, J. (2002) Positive deviance: a new paradigm for addressing today's problems today, *Journal of Corporate Citizenship, 5, pp.57-62.*

Steinke, J., Mgimiloko, M.G., Graef, F., Hammond, J., van Wijk, M.T. & van Etten, J., 2019. Prioritizing options for multi-objective agricultural development through the positive deviance approach, *PloS one*, *14*(2), e0212926.

Tucker, C.J. & Sellers, P.J., (1986) Satellite remote sensing of primary production, *International Journal of Remote Sensing*, *7*(11), pp.1395-1416.

Wishik, S. M., & Van Der Vynckt, S. (1976) The use of nutritional "positive deviants" to identify approaches for modification of dietary practices, *American Journal of Public Health*, 66(1), 38–42.

Appendix

Appendix A: Agricultural Census

Variable Code	Variable Name	Variable Values	Variable Labels
prop	Province		
kab	District		
kec	Subdistrict		
desa	Village		
Idruta	Household number		

r103	Age of head of household		
r104	Gender of Head of Household		
r201	Rice Farming business	1	Yes
		0	No
r213l	No. of Male Household managing agribusiness	Numeric	
r213p	No. of Female Household managing agribusiness	Numeric	
r214	Main types of household business	201	Rice Farming
		202	Other Crops
		203	Horticulture
		204	Plantation
		205	Livestock
		206	Fish farming
		207	Catching Fish
		208	Aquaculture
		209	Wild Animals
		2010	Agricultural Service
r217	Sex of the main famer of the main business household	1	Male
		2	Female
r301ak1_2	Code of the paddy rice plant	1101	wetland rice
r301ak2 wetland	Season1	numeric	area of rice m square
r301ak3	Season2	numeric	area of rice m square
r301ak4	Season3	numeric	area of rice m square
r301ak5	Sum	numeric	area sum
r301ak6	Main harvesting method	1	Harvested young
		2	Harvested other forms

		3	Harvested yourself
		4	Released
		5	Allowed
		6	Not harvested
r301ak7	The yields are sold/exchanged for wetland rice	1	Yes
		2	No
r301ak8	Management status	1	Manager Owned
		2	Managed with revenue sharing
		3	Managing and you are receiving a wage
r301bk1_2	Code of the rice field crops	1102	dryland
r301bk2 Rice Fields	Season1	numeric	area of rice m square
r301bk3	Season2	numeric	area of rice m square
r301bk4	Season3	numeric	area of rice m square
r301bk5	Sum	numeric	area sum
r301bk6	Main harvesting method	1	Harvested young
		2	Harvested other forms
		3	Harvested yourself
		4	Released
		5	Allowed
		6	Not harvested
r301bk7	The yields are sold/exchanged for field rice	1	Yes
		2	No
r301bk8	Management status	1	Manager Owned
		2	Managed with revenue sharing

		3	Managing and you are receiving a wage
r302	Types of rice plants that have highest production value	1101	wetland rice
		1102	Rice fields
r306a	Household members doing agricultural services other than farm labour	1	Yes
		2	No
r306b1	Household members doing rice products	1	Yes
		2	No
r901a1k2	Area of irrigated rice fields	numeric	area
	Location of irrigated rice fields	1	in the village
		2	outside the village within the subdistrict
		3	outside the sub district within the district
		4	outside the district
r901a2k2	Area of simple irrigation	numeric	area
r901a3k2	Area of rainfed	numeric	area
r901a4k2	Area of tidal swamp	numeric	area
r901a5k2	Area of wetland (swampy swamp)	numeric	area
r901a6k2	Agricultural land that is rice	numeric	area
r901b8k2	Agricultural land that is not rice	numeric	area
r902k2	Non-agricultural land	numeric	area
r903k2	Total land (Agricultural and non-agricultural)	numeric	area

Appendix B: Village Potential Survey (PODIS)

Variable Code	Variable Name	Variable Values	Variable Labels
R101	Province Code		
R102	Regency Code		
Q103	District Code		
Q104	Village Code		
R101N	Province Name		
R102N	Regency Name		
R103N	District Name		
R104N	Village Name		
R301	Government Status	1	Village
		2	Village
		3	UPT/SPT
		4	Other
		5	Natagara
R302	Consultative Body	1	Yes
		0	No
R303	Village Boundaries Lawful Map	1	Yes
		2	No
R304a	Existence of a local Environmental Unit	1	Yes
		2	No
R305B	Topography	1	Slope/Peak
		2	Valley
		3	Plain
R306	Village Head Office Location	1	Yes, in the village
		2	Yes, outside the village
		3	No Office
R307a	Village direct access to the ocean	1	Yes

		2	No
R307B1A	Fishing	1	Yes
		2	No
R307B1B	Utilization of fishing for aquaculture	3	Yes
		4	No
R307B1C	Utilization of fishing for salt ponds	5	Yes
		6	No
R307B1D	Utilization of Ocean for tourism	7	Yes
		8	No
R307B1E	Utilization of Ocean for public transportation	1	Yes
		2	No
R307B2	Existence of Mangroves	1	Yes
		2	No
R308A	Where is the village located	1	In the forest
		2	At the edge of the forest
		3	Outside the forest
R308B	Forest Function	1	Conservation
		2	Production
R4031	Residents working abroad	1	Yes
		2	No
		3	I Don't Know
R403B1	Male Migrant workers	numeric	
R403B2	Female Migrant workers	numeric	
R404A	Main source of Income	1	Agriculture
		2	Mining and excavation
		3	Manufacturing
		4	Trading
		5	Transportation

		6	Service
R404B1	Main commodity	1	Rice
		2	Other Crops
		3	Horticulture
		4	Plantation
		5	Animal Husbandry
		6	Capture fisheries
		7	Aquaculture
		8	Forestry
R404B2	Road Surface type from village to agricultural area	1	Concrete
		2	Hardened
		3	Land
		4	Other
R501A1	Families with PLN Electricity	numeric	
R501A2	Families without PLN Electricity	numeric	
R501B	Families without Electricity	numeric	
R503	Cooking fuel used by household	1	City gas
		2	LPG
		3	Kerosene
		4	Firewood
		5	Other
R504	Access to bathrooms	1	Independently
		2	Together
		3	Public toilets
		4	No toilets
R506	Drainage system	1	Infiltration hole
		2	Drainage Sewage system
		3	River or Ocean

		4	In a hole
		5	other
R507B	Source of drinking water	1	Bottle of water
		2	Plumbing with ammeter
		3	Plumbing without ammeter
		4	Drilling well
		5	well
		6	spring
		7	River or Lake
		8	Rainwater
		9	Other
R508AK2	Existence of river	1	yes
		2	No
R508AK3	Existence of Irrigation Channels	1	yes
		2	No
R508AK4	Existence of lake or reservoir	1	yes
		2	No
R508B3K2	Use rivers for irrigation of agricultural land	1	yes
		2	No
R508B3K3	Use of irrigation channels for irrigation	1	yes
		2	No
R508B3K4	Use of lakes and reservoirs for irrigation	1	yes
		2	No
R511A	Existing of Slums	1	yes
		2	No
R512AK2	Pollution Incident (water)	1	yes
		2	No
R512BK2	Pollution Incident (Soil)	1	yes

		2	No
R512CK2	Pollution Incident (Air)	1	yes
		2	No
R513	Burning fields for agricultural purposes	1	yes
		2	No
R601AK7	How many landslides in 2013	numeric	
R601BK7	How many flood in 2013	numeric	
R601CK7	How many Flash Flood in2013	numeric	
R601DK7	How many earthquakes in 2013	numeric	
R601JK7	How many drought events in 2013	numeric	
R701AK2	Number of levels of education	numeric	
R702A	Functional Literacy activities	1	Yes
		2	No
R704AK2	Hospital Facilities	1	Yes
		2	No
R709(All)*	Health Epidemics	1	Yes
		2	No
R710	Number of residents suffering from bad nutrition	numeric	
R807A	Habit of mutual cooperation of residents	1	Yes
		2	No
R1001B2	Roads can be accessed by cars or larger	1	All year long
		2	all year except certain times
		3	all year except wet season

		4	Not passable
R1103A_K4	Conversion from rice to non-rice agriculture	1	Yes
		2	No
R1103C_K2	Conversion from non-rice to rice agriculture	1	Yes
		2	No
R1205	Number of markets without buildings	numeric	
R1212B	Number of small industry cooperatives	numeric	
R1212C	Number of savings and loans cooperatives	numeric	
R1213A	The existence of stalls that sell agricultural production facilities owned KUD	1	Yes
		2	No
R1214C	Small Business credit facility received by residents	1	Yes
		2	No
R1401A4_K2	Programmes community development, irrigation, markets, agriculture	1	Yes
		2	No
R1401A4_K3	Source of programme intervention	1	PNPM
		2	Non PNPM
		3	PNPM and Non PNPM
R1401A4_K4	Programme implementers	1	poor population
		2	non-resident
		4	farmer
		8	business group

		16	other
R1401A4_K5	Direct beneficiaries	1	poor population
		2	non-resident
		4	farmer
		8	business group
		16	other
R1401B1_K2	Programmes for capacity building, lending for agriculture	1	Yes
		2	No
R1401B1_K3	source of the programme	1	PNPM
		2	Non PNPM
		3	PNPM and Non PNPM
R1401B1_K4	Implementers	1	poor population
		2	non-resident
		4	farmer
		8	business group
		16	other
R1401B1_K5	Direct beneficiaries	1	poor population
		2	non-resident
		4	farmer
		8	business group
		16	other
R1501A_K3	Revenue	numeric	
R1501B_K2	Village fund allocation value	numeric	
R1503C	The existence of village market assets	5	Yes
		6	Nothing
R1601A_K	Gender of the head	1	Yes
		2	Nothing
R1601A_K5	Education of the head	1	Never attended school

		2	Not finished
		3	Graduated from elementary school / equivalent
		4	Junior high school / equivalent
		5	High school / equivalent
		6	Academy / DIII
		7	Diploma IV / S1
		8	S2
		9	S3
KCR803B2K2	Number of markets specifically fruit and vegetables	numeric	
KCR803B3K2	Number of special markets for rice	numeric	
KCR803B3K3	Special types of rice market buildings	1	Permanent
		2	Semi-permanent
		4	No Building
KBR801A	Disaster management efforts	1	Yes
		2	Nothing

Chapter Five: Data-powered positive deviance: Combining traditional and non-traditional data to identify and characterise development-related outperformer

Basma Albanna, Richard Heeks, Andreas Pawelke, Jeremy Boy, Julia Handl and Andreas Gluecker

Abstract

The positive deviance approach in international development scales practices and strategies of positively-deviant individuals and groups: those who are able to achieve significantly better development outcomes than their peers despite having similar resources and challenges. This approach relies mainly on traditional data sources (e.g. surveys and interviews) for identifying those positive deviants and for discovering their successful solutions. The growing availability of non-traditional digital data (e.g. from remote sensing and mobile phones) relating to individuals, communities and spaces enables data innovation opportunities for positive deviance. Such datasets can identify deviance at geographic and temporal scales that were not possible before. But guidance is needed on how this new data can be employed in the positive deviance approach, and how it can be combined with more traditional data to gain deeper, more meaningful, and context-aware insights.This paper presents such guidance through a data-powered method that combines both traditional and non-traditional data to identify and understand positive deviance in new ways and domains. This method has been developed iteratively through six development projects covering five different domains – sustainable cattle ranching, agricultural productivity, rangeland management, research performance, crime control – with global and local development partners in six countries. The projects combine different types of non-traditional data with official statistics, administrative data and interviews. Here, we describe a structured method for data-powered positive deviance developed from the experience of these projects, and we reflect on lessons learned. We hope to encourage and guide greater use of this new method; enabling development practitioners to make more effective use of the non-traditional digital datasets that are increasingly available.

5.1 Introduction

Positive deviance (PD) is based on the observation that in every community or organisation, a few individuals or groups develop uncommon practices or behaviours to produce better solutions to problems than their peers who face the same challenges and barriers (Pascale, Sternin & Sternin 2010). Those individuals are referred to as positive deviants and adopting their solutions is referred to as the PD approach. This is an approach that, particularly since the turn of the century, has found a growing niche within development research and practice. However, there are challenges that have constrained the spread of the PD approach; some of which are data-related. Recognising this, it has been proposed that recent developments in the increasing availability of non-traditional digital data provides an opportunity to identify and understand positive deviants in new ways; potentially helping address some of these challenges (Albanna & Heeks 2019). We refer to the use of such non-traditional data to replace or complement traditional data as the "data-powered positive deviance" (DPPD) method. 'Non-traditional data' in this context broadly refers to data that is digitally captured (e.g. mobile phone records and financial data), mediated (e.g. social media and online data) or observed (e.g. satellite imagery). 'Traditional data' refers to data captured manually such as official statistics, observation data, surveys and interviews.This paper provides an exposition of the DPPD method by describing a methodological framework that guides the combined use of traditional data sources and non-traditional digital data sources to identify and characterise positive deviants in development-related challenges. The framework was first outlined by Albanna and Heeks (2019) and then further developed through its application in a global initiative collaboratively conducted by the Deutsche Gesellschaft für Internationale Zusammenarbeit (GIZ) Data Lab, United Nations (UN) Global Pulse Lab Jakarta, the United Nations Development Programme (UNDP) Accelerator Labs Network and the University of Manchester (Data Powered Positive Deviance 2020). This initiative refined the DPPD method by applying it to five distinct domains, spanning six developing countries to identify and understand: farmers achieving higher than usual cereal crop productivity in Niger and Indonesia; cattle farmers in Ecuador who are deforesting below average rates; research output outperformance among Egyptian researchers; public spaces in Mexico City where women are safer; and communities in Somalia which are able to preserve their rangelands despite frequent droughts. The framework presented here should provide a tool for development

professionals to identify outperformance in different development sectors by mixing analytical insights from traditional and non-traditional data. Such insights should help amplify innovative, locally-sourced and evidence-informed solutions to development challenges.

In what follows, we first present the history and challenges of positive deviance and the potential for non-traditional data and data science to address those challenges. Following that, we explain how we developed the DPPD method. We then present the three core stages of the method: assessing problem-method fit, determining positive deviants, and discovering the underlying factors leading to positive deviance. We summarise preliminary results from applying these stages in the pilot projects, then discuss lessons learned from applying DPPD, and end with conclusions, including thoughts on future application of the method.

5.2 Background

Positive deviance was used for the first time in 1976 to inform the design of food supplementation programmes in Central America by identifying dietary practices developed by mothers in low-income families having well-nourished children (Wishik & Van Der Vynckt 1976). The full method and results of this study were not published, limiting uptake. However, in the 1990s the PD approach became more widely recognised as a credible strategy for operational and academic research in nutrition, based on extensive observations and a strong emphasis on impact (Zeitlin 1991; Sternin et al. 1997; Sternin et al. 1998). Its first large-scale adoption was by Save the Children Foundation, which used PD as a strategy to reduce malnutrition in Vietnam, rehabilitating an estimated 50,000 malnourished children in 250 communities (Sternin 2002). But it was not until the 2000s that PD started to attract wider attention, when Sternin and colleagues introduced it as a development approach for social change and demonstrated how it can be operationalised as a domain-agnostic approach (Sternin & Choo 2000; Sternin 2002). Since then, PD has been applied across multiple development domains, with public health being the most prominent.

Originally, the PD approach was designed to study the characteristics and practices of individuals who are able to achieve better results in response to a specific development challenge. A more recent set of PD studies has been interested in how certain individuals or

groups respond to a development intervention programme significantly better than their peers who are targeted by the same intervention (Post & Geldmann 2018). This is similar to randomised control trials in the sense that it compares post-intervention performance with pre-intervention performance. But in PD studies, the interest is not in the difference in performance between the control and intervention groups as much as in the variation in the performance of units within the intervention group, and potential factors that led to this variation. Identifying the reasons behind the exceptional response of the positive deviants can be used to inform intervention strategies and to increase overall adoption by "bad responders".

Notwithstanding the growth in prevalence of positive deviance as an approach to international development, its adoption has been constrained by a number of challenges (Albanna & Heeks 2019). Given these challenges, there are obvious opportunities for innovation in PD and our particular interest here is in the innovative opportunities offered by non-traditional, digital data sources like big data following the increasing "datafication" of development and growing availability of big datasets in a variety of development sectors (Hilbert 2016). The opportunities have been identified via a systematic literature review of positive deviance and big data in development (Albanna & Heeks 2019):

- **Time and cost of data collection:** traditional PD studies rely mainly on primary data collection to identify positive deviants and to understand their underlying practices and strategies; something that involves significant time, cost and risk. These could be ameliorated by use instead of existing big datasets if they contain indicators of relevance to positively-deviant performance.
- **Positive deviant identification:** because of the costs of data collection, and notwithstanding the substantial impacts of some PD projects, the overall population sample in traditional PD studies tends to be relatively small[35]. Given they are the exceptions in any population, this makes the number of positive deviants in these

[35] The sample size is typically tens or hundreds at most, with the exception of a few studies in health care (Bradley et al. 2009; Wallace &Harville 2012).

samples very small, constraining generalisability of conclusions about their particular features and practices. Big data, by contrast, may cover large populations, making it possible to identify a larger number of positive deviants and thus to improve the generalisability of conclusions for practice. Additionally, data sources in traditional PD studies provide a static, cross-sectional reflection of performance whereas some big data could provide a dynamic picture due to its longitudinal coverage. Finally, traditional PD studies have tended to focus on individuals or individual households as positive deviants because they are most amenable to field survey methods. Big data might offer opportunities via direct coverage or aggregation to identify positively-deviant communities or even regions.

- **Monitoring and evaluation:** because of time, cost, logistical and other challenges, traditional PD studies rarely evaluate the impact of any interventions developed as the result of positive deviant identification and analysis. If a big dataset longitudinally captures relevant performance indicators, then it could relatively easily be used for monitoring and evaluation of the effects of scaling positively-deviant practices into an intervention population.
- **Expanding the scope of PD:** despite the spread of positive deviance noted above, there has been a domain and geographic skew in its application. According to Albanna and Heeks (2019), 89% of the sample of PD studies they reviewed were in public health (a form of path dependency due to the success of its first application in nutrition), with 83% targeting rural communities. Big data could help break PD from its current narrow focus, due to the existence of big datasets dealing with a variety of development domains and locations.

However, the role of big data in development has itself been criticised, given that big datasets may often be decontextualised (Taylor & Broeders, 2015). A number of studies have suggested integrating "thick data"[36] with big data to extract meaning and value from it and to rescue it from the potential context loss (Bornakke &Due, 2018; Smets & Lievens, 2018; Ang 2019). Such

[36] Data collected through qualitative and ethnographic methods to uncover individual behaviours and attitudes (Bornakke & Due 2018).

a combination could be seen as particularly relevant for positive deviance. In order to identify "true" positive deviants[37] that are performing unexpectedly well due to uncommon behaviours and strategies, it is crucial to control for the contextual variables that could influence this performance. Given its decontextualised nature, big data rarely contains such variables and they must therefore be sought in other, traditional data sources. Therefore, combining traditional data with big data is an integral part of the DPPD method presented in this paper.

5.3 Methodology

The potential value of non-traditional data to the positive deviance approach can only be realised if action researchers and practitioners are provided with a clear method through which to make use of these data. The aim of this paper is therefore to present a systematic method for data-powered positive deviance (DPPD) by testing and validating the use of big data and other types of non-traditional data in positive deviance. In order to do this, we built on the preliminary framework proposed by Albanna and Heeks (2019) which sought to integrate the use of big data into the five main stages of the PD approach (Positive Deviance Initiative 2010):

"1) *Define* the problem, current perceived causes, challenges and constraints, common practices, and desired outcomes.
2) *Determine* the presence of positive deviant individuals or groups in the community.
3) *Discover* uncommon but successful practices and strategies through inquiry and observation.
4) *Design* and implement interventions to disseminate PD practices and strategies.
5) *Monitor* and evaluate the resulting project or initiative".

The first version of the DPPD method was applied in a case study of Egyptian researchers who outperformed their peers in terms of research outputs. Following that, we iteratively

[37] Positive deviants who are not false positives that were mistakenly identified as positive deviants because of a contextual advantage that was not accounted/controlled for.

developed the method through a collaborative initiative between the GIZ Data Lab, UN Global Pulse Lab Jakarta, the UNDP Accelerator Labs Network and the University of Manchester. Action research was chosen as the research strategy because it bridges the gap between research and practice by integrating, rather than chronologically separating, the two processes of research and action (Somekh 1995). It would therefore allow the application of the DPPD method to be fed back into its conceptualisation; that re-conceptualisation then refining practice in an iterative cycle.

In addition to the Egypt case study, the action research cycles were applied in five other PD projects (see Table 27). These were chosen following a call for proposals, to which GIZ field offices and the UNDP Accelerator Lab Network responded. Proposals were selected based on judgement of their viability and the diversity of development domains and countries to which the DPPD method could be applied. As shown in Table 1, the projects also offered diversity in terms of non-traditional data types – citation data, remote sensing data, mapping and cadastral geographic data – and both proprietary and open data sources. This was complemented by a variety of traditional data sources: official statistics, administrative data, surveys and interviews. The units of analysis covered different aggregation levels starting with individuals, farms and communities up to geographical units representing urban areas and villages. This diversity of domains, countries, data and scales was seen as important in helping to broaden the testing base for the DPPD method and to strengthen its likely generalisability.

Within the overall collaborative initiative, a central group was responsible for revising the DPPD method. Its application was led by country-level practitioner teams drawn from domain specialists in GIZ field offices and UNDP Accelerator Labs working in continuous contact with the central group.

The DPPD method that emerged from this process follows the same five stages as the PD approach outlined above, but uses pre-existing non-traditional data sources instead of – or in conjunction with – traditional data sources across the five stages. As detailed in the following section, this requires a series of new and specific methods and practices that are not required in the conventional PD approach. The first stage is also somewhat different because it not only defines the problems but also checks if it is suitable and feasible to use the DPPD method for the proposed project.

Project	Unit of Analysis	Definition of Positive Deviants	Data Used
Research publication outperformance in Egypt(Albanna, Handl & Heeks 2021)	Individual researcher	Information systems researchers in public universities who achieved significantly higher-than-average scores in one or more of six citation metrics	Citation data from Google Scholar, research publications on Scopus, university websites, interviews and surveys
Rice-farming outperformance in Indonesia (Albanna, Dhar Burra & Dyer 2020)	Village	Rice-farming villages that have higher than expected rice productivity as measured by Enhanced Vegetation Index (EVI) scores while controlling for their climatic, socio-economic and demographic conditions	Remote sensing data, official statistics, administrative boundary data and crop masks
Rangeland preservation by pastoral communities in Somalia (Abdullahi, Albanna & Barvels 2021)	Community	Communities in the same land capability class that were able to sustain or enhance their rangelands' health (since the 2016 drought) as measured by the Soil-Adjusted Vegetation Index (SAVI)	Remote sensing data, settlement location data, observation data and semi-structured interviews
Cereal-farming outperformance in Niger (Gluecker, Lehman & Barvels 2021)	Community	Communities which – despite drought and conflict – achieve high cereal yields as calculated by higher than expected SAVI while controlling for soil, evapotranspiration, precipitation and land use	Remote sensing data, administrative boundary data, land use data, observation data and semi-structured interviews

Public spaces in Mexico City where women are safer (Cervantes, Rios & Soto 2021)	AGEB[38]	AGEBs where gender-based crimes and crimes with female victims are lower than expected given their population density, demographics, socio-economic status and urban infrastructure	Mexico City open data portal, 911 calls, administrative boundary data, official statistics, observation data and semi-structured interviews
Low deforestation cattle farming in the Ecuadorian Amazon (Grijalva et al. 2021)	Farm	Cattle-raising farms operating in areas of potential forest clearance with deforestation rates that are significantly lower than expected for three consecutive years, while controlling for the size of the farm, the land use, soil adaptability, socio-economic status and cattle density	Remote sensing data, vaccination data, cadastral data, official statistics, land use data, observation data and semi-structured interviews

Table 27: Summary of DPPD method projects

5.4 The Data-Powered Positive Deviance Method

This section presents the three core stages of the DPPD method. We focus on these three for two reasons: first, because these are the stages so far achieved by the action research projects and, second, because these are the stages that differ most significantly from the conventional PD approach and which therefore most require new guidance. Stage 1 defines the problem and validates if it is suitable and viable to use the DPPD method, hence we refer to it as 'Assessing problem-method fit' instead of 'Defining the problem' (the original name of this stage in the PD approach). Stage 2 seeks to identify positive deviants within the available datasets, and Stage 3 seeks to uncover the factors underlying positive deviant

[38] Área Geoestadística Básica (AGEB), which is the basic geo-statistical area in Mexico City.

outperformance. Figure 33 provides a summary of these three core stages of the DPPD method and the different steps conducted in each stage.

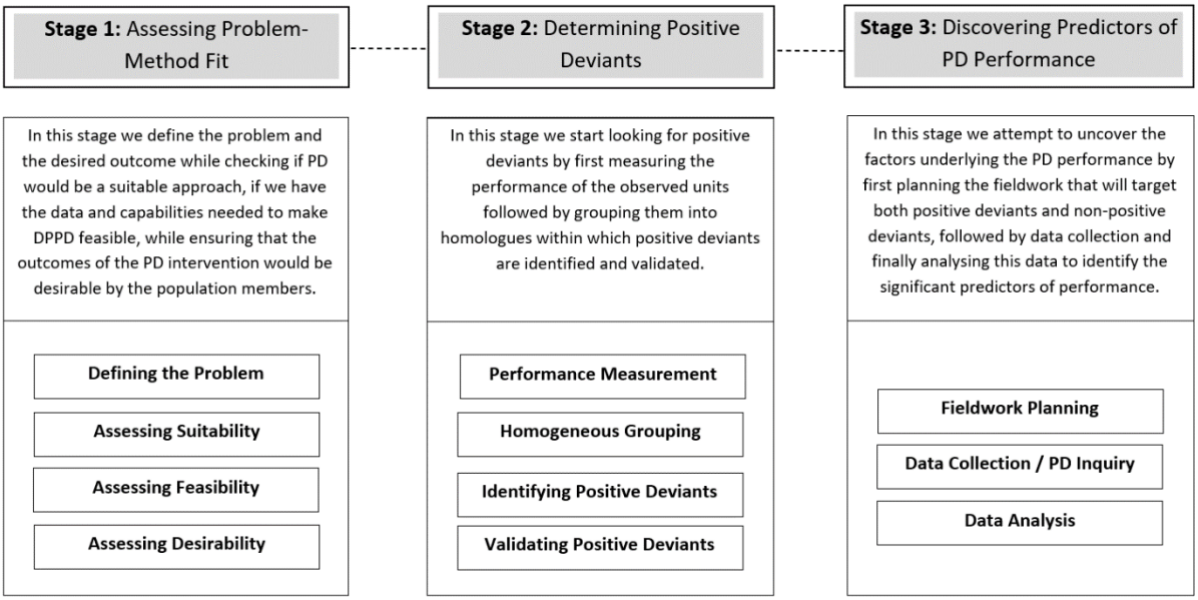

Figure 33: The first three stages of the DPPD method

5.4.1 Stage 1: Assessing Problem-Method Fit

In a similar way to the positive deviance approach, the first step of the data-powered positive deviance method is to define the problem and the desired outcome. However, in order to move from the problem to the desired outcome, one has to make sure that using a PD approach is suitable for the problem at hand, to check access to the various data sources and capabilities needed to identify and characterise positive deviants, while ensuring no harm is likely to affect the observed units. So, before applying the DPPD method to the identified problem, it is important to first answer three main questions:

- **Suitability**: Is the positive deviance approach suitable to address this type of development problem?
- **Feasibility**: Is there access to data sources and capabilities that would make it feasible to reach the desired outcome using the DPPD method?
- **Desirability**: Who is likely to benefit from or be harmed by the project, including any potential unintended negative consequences from data analysis?

Defining the Problem

When defining the problem, it is important to specify the study population and the unit of analysis. The 'study population' is the group of individuals, communities or geographic units who are suffering from or causing the problem and will be included in the analysis. The 'unit of analysis' is the level at which one can find solutions to the addressed problem. For example, in the Mexico safe public spaces project, the problem we are trying to tackle is the high rates of violence against women in public spaces. The study population is public spaces in Mexico City, our units of analysis are AGEBs and the desired outcome is to reduce violence against women and girls in public spaces (Cervantes, Ríos & Soto 2021). In this step, it is also important to identify the different stakeholders who should be involved (community members and leaders, development professionals, government officials, etc.) in discussions around the current perceived causes of the problem, and to better understand the community's context including challenges and constraints, existing human and natural resources, common practices and normative behaviours. Having the buy-in of the different stakeholders at the very beginning guarantees, to some extent, the adoption and amplification of findings from the PD inquiry later on.

Assessing Suitability

There are two key criteria to determine whether a PD approach is suitable: 1) The nature of the development problem being addressed, and 2) The likelihood that positive deviants exist. Neither the conventional PD approach nor DPPD will be suitable if the addressed problem requires mainly a technical solution, e.g. building a road or constructing a dam – in such circumstances, the positive outcome is likely not related to individual practices and strategies. A PD / DPPD approach is much more likely to be applicable if the problem has social components and requires some form of behavioural change or a shift in mindsets, as seen in the project examples discussed here.

Even in this situation, before starting a PD intervention, it is important to check if positive deviants exist. While it may be hard to do this before diving into the data, there are ways to assess whether positive deviants exist or not by engaging with relevant stakeholders that are concerned with the issue at hand. Meeting with key development actors including

community leaders and government officials familiar with the targeted sector will help give a sense if outperformers exist. It is also useful to review academic and grey literature to check if there are direct examples of positive deviants or indirect evidence of local solutions sourced from communities living in comparable contexts and facing similar challenges. For example, before starting the Ecuador cattle-farming project, we knew through conversations with a key development actor that certain farmers adopt more sustainable cattle-ranching practices and deforested less than others. Similarly, in Somalia, through interviews with an officer from the Ministry of Environment and Rural Development, we learned about a positively-deviant community that was protecting its trees from cutting and burning for charcoal production. Figure 34 shows four PD suitability quadrants that can be used to judge whether a PD approach is suitable or not. The projects for which the PD approach is best suited generally lie in the top right quadrant, where positive deviants are likely to exist, and scaling their practices should contribute to solving the problem. Projects in the bottom left quadrant, where positive deviants are unlikely to exist, and scaling practices will likely have a limited impact on the problem, are not suitable for the PD approach. Some further consideration is given to the issue of suitability in Box 5.1.

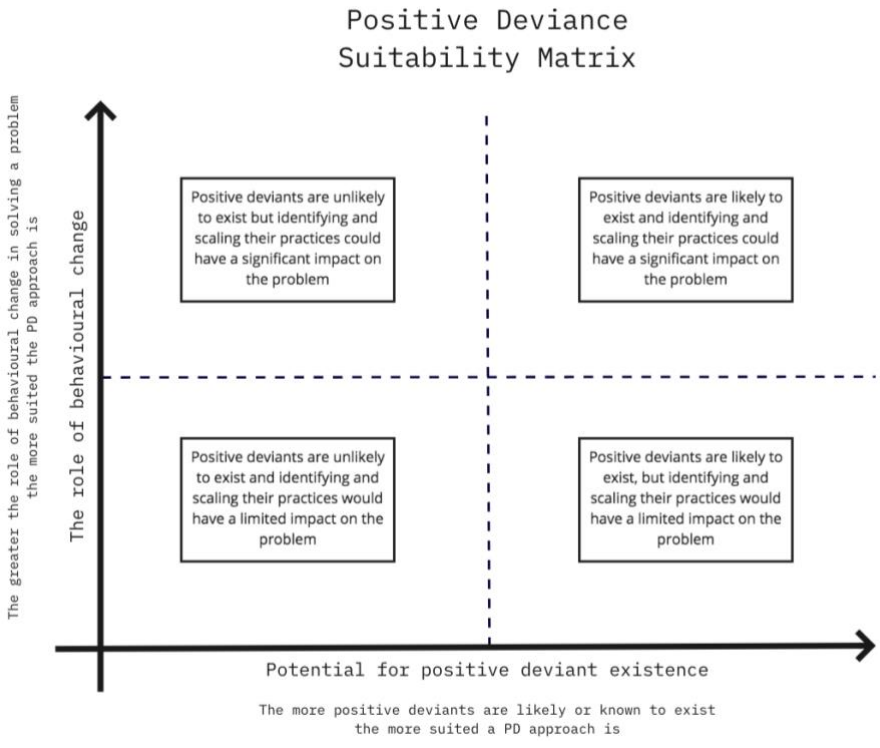

Figure 34: Positive deviance suitability matrix

Box 5.1 Further Potential Dimensions of PD Suitability Determination

While this did not arise explicitly within the projects, it may be possible to hypothesise potential other characteristics of the problem landscape, beyond those in Figure 34, that lend themselves for or against application of PD. We did this by asking ourselves what are areas where PD has more chances of success, and where could some red flags show that it might not be the right solution for a problem?

In a number of the pilot projects, we encountered the "tragedy of the commons" i.e. situations were individuals with access to a common or shared resource act in their own interest and, contrary to the common good of other individuals in their community, causing a depletion of the resource due to uncoordinated action (Hardin 1968). Case in point, the use of rangelands by pastoralists in the Somalia project or the frontier logging of the Ecuadorian Amazon forest by cattle farmers in the Ecuador project. Before lending such problems to the DPPD method, it is important to ensure that solutions mitigating this tragedy can be found. Elinor Ostrom (1990) identified criteria in the selection of cases where it could be possible to overcome the tragedy of the commons. These criteria could also be used to guide the selection of PD projects before falling prey to the "tragedy of the commons". One criterion could be to ensure that resources have **confined boundaries**, **can be preserved, and are scarce**. The latter would warrant that the users will have **strong incentives** to manage their resources in a sustainable fashion. Another important criterion is to investigate the presence of communities with a thick social network and social norms that **promote conservation**. Additionally, there should be **resource economic dependence** which would guarantee that failures by resource appropriators cannot be attributed to economic indifference (Ostrom 1990).

Another dimension to determine PD suitability is related to the economic concept of "rivalry" (Mankiw 2012). A good is said to be **rivalrous** if its consumption by one consumer prevents simultaneous consumption by other consumers. On the other hand, a good is considered **non-rivalrous**, if for any level of production, the cost of providing it to a marginal individual is zero. By adding excludability (i.e. the degree to which a good, service or resource can be limited to only paying customers) to rivalry, four types of consumption goods were identified: Private Goods, Common-Pool Goods, Club Goods and Public Goods, (Mankiw 2012). **Private**

goods that are both rival and excludable are the least suitable for positive deviance because the providers of the goods might feel that they will lose a competitive advantage by exposing their strategies to others e.g. a store/seller achieving significantly higher sales in a specific product. However, if the production of such goods is for self-consumption, the application of positive deviance could be suitable. For instance in the Niger agricultural pilot, most of the farmers grew cereals for food self-sufficiency, hence they did not feel the threat of losing a market advantage by exposing their strategies. **Common-pool goods** that are rival and non-excludable (e.g. fish in the ocean) can be linked to the tragedy of the commons and here we can follow the aforementioned criteria in picking potential PD projects where solutions can be found and scaled. **Club goods** that are non-rival but excludable (e.g. subscription-based websites), similar to private goods, are not very well suited for positive deviance since competition would still be there between providers of goods due to its excludability. However, identifying and understanding positive deviance in the usage of the goods could be feasible e.g. content achieving the highest viewership. In the case of development, it could be adopters of a new technology for better agricultural outcomes such as drought-resilient crop seeds. Finally, a situation with **public goods** that are both non-rival and non-excludable (e.g. public safety, environment) could lend itself to PD research if viewed from the lens of how users interact with it, and which interactions lead to the desired outcomes and why. For instance, public spaces that are safer for women or areas where biodiversity conservation efforts are successful.

Assessing Feasibility

The DPPD method relies heavily on existing non-traditional digital datasets that complement more traditional secondary data to identify positive deviants. Given the dependency on existing datasets for the DPPD method to work, a number of conditions regarding data availability, accessibility and adequacy need to be met. In terms of **availability**, it is important to ensure that there are outcome and contextual data, which are crucial for the identification of positive deviants. Outcome data is used to directly or indirectly measure the performance of the target group. It should be capable of identifying individuals, groups or, more generally, units within the target group that outperform their peers. DPPD leverages the potential of readily available non-traditional, digital data, like earth observation data, online or mobile data, citizen-generated data, or sensor data, to capture outcomes[39]. Contextual data is then needed to control for factors that are likely to impact performance but are not related to practices or behaviours. A large portion of those factors should emerge from the perceived causes of the problem identified in the previous step. Contextual data helps put the outcome measure in perspective and guarantees that positive deviants are identified based on performance relative to comparable peers rather than absolute performance. This data can be extracted from traditional data sources (e.g. census) and non-traditional data sources (e.g. remotely sensed climate data). Both the outcome and contextual data should be spatially and temporally relevant to the problem at hand, i.e. data should be at an aggregation level that is sufficient to capture the outcomes of the unit of interest and it should be recent enough to ensure the relevance of the field investigation.

For instance, in the Niger and Indonesia projects, when measuring the agricultural outcomes of cereal farming villages, we needed remote sensing-based measures of vegetation health. To extract those village-delimited vegetation indices, administrative data about their boundaries was required. And to identify positive deviants within groups sharing similar resources and context, we needed digitally available climate and soil data, administrative data

[39] More detail on this is provided in section 5.4.2.

about land cover and agroecological zones, and official statistics in order to capture their socio-economic and demographic conditions (Albanna et al. 2020; Gluecker et al. 2021).

After identifying available relevant data, comes the question of data **accessibility**. When there is a need to use non-public data, having data access agreements in place with data providers is a real asset. It can take significant time to negotiate the conditions over access to such data. The Ecuador cattle-farming project grew out from a current project called ProAmazonia, run by UNDP and the Ministries of Environment and Agriculture. Through this partnership it was possible to access cattle vaccination data from the Ministry of Agriculture, training datasets for land cover analysis through the Ministry of Environment, and cadastral data as well as farm boundary data from municipalities through ProAmazonia.

Finally, and most importantly, the available and accessible data should fit the scope of the project, the know-how needed for the analysis should be attainable, and the choices of both data and skills should account for the project's time and budget limitations. This data **adequacy** is usually achieved after several iterations between problem framing and data mapping until a suitable match is found. This process should not compromise the initial purpose of the project. It might however call for starting with a somewhat flexible problem definition (or lens) in well-defined domains with clear development challenges. This flexibility allows navigation through different proxy options to capture the outcomes of the target group. In the Somalia rangelands project, we moved from identifying drought resilient pastoral communities by measuring their livestock numbers remotely, to identifying them based on their ability to sustain the health of their rangelands, which is necessary to maintain livestock (Abdullahi, Albanna & Barvels 2021). The data for the former was costly to obtain (very high resolution imagery) and required extensive and complex analytical skills that are rare to find, whereas the data for the latter was readily available and could be more easily analysed because the team already has a remote sensing analyst at hand.

The process of "Assessing Feasibility" presented in this section may appear linear. In practice, though, an accessible data source may turn out to be not available, or an available source may transpire to be not adequate. Hence, the actual process will be iterative until data is found that enables reliable capture of the outcomes and context of the target group.

Assessing Desirability

This step is about closely looking at all those individuals and communities that stand to benefit or lose from the identification and amplification of the practices, strategies and other factors associated with the outperformance of positive deviants. Should this assessment yield a potential outcome that would harm those three groups – the positive deviants, the non-positive deviants, or the wider community – it may be advisable to adjust the overall design of the project or to abandon the idea. Having said this, some evaluation of net public benefit needs to be undertaken. Some reordering of incentives and benefits may be appropriate, even if it disadvantages one group, if it is to the greater public good.

The PD approach assumes that positive deviants are not aware of their innovativeness and/or impact of their uncommon practices and strategies. But what if they are aware and have deliberately chosen not to share their strategies with other members of the community? For example, they might fear losing their competitive advantage over others, or depleting a resource they alone are aware of, if it were to be shared with other community members, which makes this solution unsustainable. Hence, it is important to assess if it is desirable for a positive deviant to share their practices and behaviours with others. This is likely to be less of an issue where cultural norms dictate against competitive strategies, such as child malnutrition or health. However, it might be more problematic in areas where people more overtly compete with one another, as in the Egypt research performance case study. The following questions can be used to assess desirability: Is it generally safe to assume that it will be desirable to scale the behavioural practice in question? Are we endangering the competitive advantage of positive deviants by sharing their practices and strategies with others? Might we risk harming a positive deviant or a non-positive deviant by revealing their identity? Will inviting others to adopt a PD strategy trigger this strategy's obsolescence? As an example, there was concern about what might happen if we promoted a particular transhumance destination in Somalia as a strategy to help community rangelands recover, given this might lead to overgrazing of rangelands at the promoted destination.

If the DPPD method is seen to be feasible from a data and capabilities point of view, it is important to ensure protection of the privacy of individuals and communities involved: "The availability or perceived publicness of data does not guarantee lack of harm, nor does it mean

that data creators consent to researchers using this data" (Zook et al. 2017). Questions to be asked here are: Is the data to be used of sensitive nature, e.g. personally-identifiable information? Are safeguards in place for safe and secure data access and processing? Has consent been given (directly or indirectly) by the data subjects to use this data? To whom can the identity of positive deviants and non-positive deviants be revealed? Given the next stage is the identification of positive deviants, such questions must be thought through at this point.

5.4.2 Stage 2: Determining Positive Deviants

After defining the problem and ensuring the applicability of the DPPD method comes the stage of looking for the positive deviants. This section outlines the different steps of this stage, starting with performance measurement, followed by homogeneous grouping and positive deviant identification, and finally the preliminary validation of the potential positive deviants.

Performance Measurement

This step attempts to identify the core performance measure for positive deviance; a measure that captures a desirable development outcome as defined by the different stakeholders of the investigated problem. The DPPD method advocates deriving this measure from non-traditional, digital data sources (e.g. big data), and using it either alone or in combination with some other measure. For instance, in the Niger and Indonesia agricultural projects, we used remotely-sensed vegetation indices (e.g. SAVI and EVI) to measure vegetation health – and, hence, agricultural productivity – as the core performance measure of agricultural communities (Albanna et al., 2020; Gluecker et al., 2021).

The data sources that are available are often collected for a different purpose than that of a positive deviance project. In such cases, it is possible that data provides only indirect insights into the subject of interest, rather than direct measures. Hence, the data source measures should be considered as proxies of the actual phenomena that need to be measured, and the validity of using these proxies must be ascertained. This validation could be as simple as checking prior literature showing a strong correlation between the proxy and the desired

outcome in a context similar to the one being investigated. If prior studies are lacking, then the proxy relationship should be ground-truthed using direct measures of performance, and their suitability should be validated with local domain experts; triangulating between multiple experts to reduce the dangers of bias within any individual source. In the Niger agricultural project, for example, use of SAVI – rather than other measures – as an indicator of vegetation health and crop productivity was based on local expert advice that SAVI was suitable for semi-arid areas like Niger given it incorporated a soil brightness correction factor (Gluecker et al. 2021).

Depending on how the desired development outcome is defined, the study can have one or multiple performance measures. In the Egypt research publication case study, six research citation metrics were used to evaluate performance because these enabled a balanced consideration of both scientific productivity and impact while controlling for factors like article and author age (Albanna et al. 2021). From each metric, positive deviants were identified and the final set of outperformers included positive deviants from all six metrics. There are also techniques that can be used to summarise multiple performance measures into a single index or at least into fewer measures. A basic approach here assumes all measures to have equal weight. If the performance measures are seen to be of varying importance, a weighted average can be used but this of course requires some collective determination of the weight (importance) of each measure. Summarisation techniques (assuming that measures have equal weight) include principal component analysis, which replaces the original set of measures with a smaller number of uncorrelated measures that account for most of the information in the original set (Abdi & Williams 2010).

Homogeneous Grouping

Having identified the measure that determines positive deviance, the next step is to divide the study population into homogeneous or peer groups having similar contextual factors, to then identify individuals who deviate positively from their peers. This way positive deviants are identified relative to their context and not in an absolute sense. This grouping also increases the likelihood of identifying positive deviance that can be attributed to particular attributes, practices and strategies that can be transferred; not deviance due to structural factors – contextual variance that impacts the studied outcome but is beyond the control of

the unit of analysis – that cannot be transferred. Figure 3 provides an illustrative example of grouping from the Mexico safe public spaces project where areas were divided into three peer groups based on socio-economic level, daily incoming trips and population density.

The grouping procedure can be done manually based on professional experience and intuition or can be done through unsupervised machine learning techniques such as clustering. The aim of the grouping is to minimise the variance of those structural factors within the groups and maximise it between the groups. There are three main drivers for this grouping:

1. The essence of the PD approach is to uncover context-aware solutions that are associated with the performance of positive deviants. Since it is difficult to capture all the contextual factors driving performance, one aim is to group observations based on aggregated structural variables (e.g. a district poverty index) that would correspondingly reduce variance from underlying disaggregated contextual factors that are not accessible (e.g. household income).
2. A number of studies (Nathan & McMahon 1990; Trivedi et al. 2011; Trivedi et al. 2015) demonstrate how clustering a population into homogeneous groups can make the per-cluster-prediction better. This is because, rather than seeking to build models that explain the natural variation between clusters, the focus is on within-cluster variation, which increases the model's performance.
3. When a study population is divided into homogeneous groups, findings can be extrapolated with more confidence (Nathan & McMahon 1990). This is because more detailed and localised information can be extracted from homogeneous groups having similar conditions (Kovács et al. 2014). This is particularly useful in the following DPPD stages that seek to uncover positive deviants' practices and strategies and to design interventions that disseminate them.

One of the main challenges associated with homogeneous grouping or clustering is the selection of the variables that will be used to assess the degree of similarity between the different observations. Clustering techniques are capable of generating clusters with literally any set of variables, so it is crucial to select variables based on their relevance to the problem. In the DPPD projects, our clustering variables were selected based on theoretical and

empirical research indicating that they have a significant impact on the outcome measure. For example, there is a well-established relationship between socio-economic conditions and crime rates (Vilalta & Muggah 2016). Therefore, it was crucial, in the Mexico safe public spaces project, to cluster urban areas into groups with similar socio-economic levels, as shown in Figure 35. It is also possible to find existing groupings of homologues or clusters that were created for a different purpose but which can be reused for PD identification. As an example, in the Niger agricultural project, we found a recently-released map developed by the Adapt'Action Facility that divided Niger into agroecological zones with similar biophysical, ecological and climatic conditions (Hauswirth et al. 2020). This zoning also included access to data on natural resources, land tenure, farming systems, socio-demographic and economic status. We decided to use these zones instead of creating our own homologues, especially as they took into account valuable local and contextual knowledge which would have been difficult for us to incorporate in our grouping.

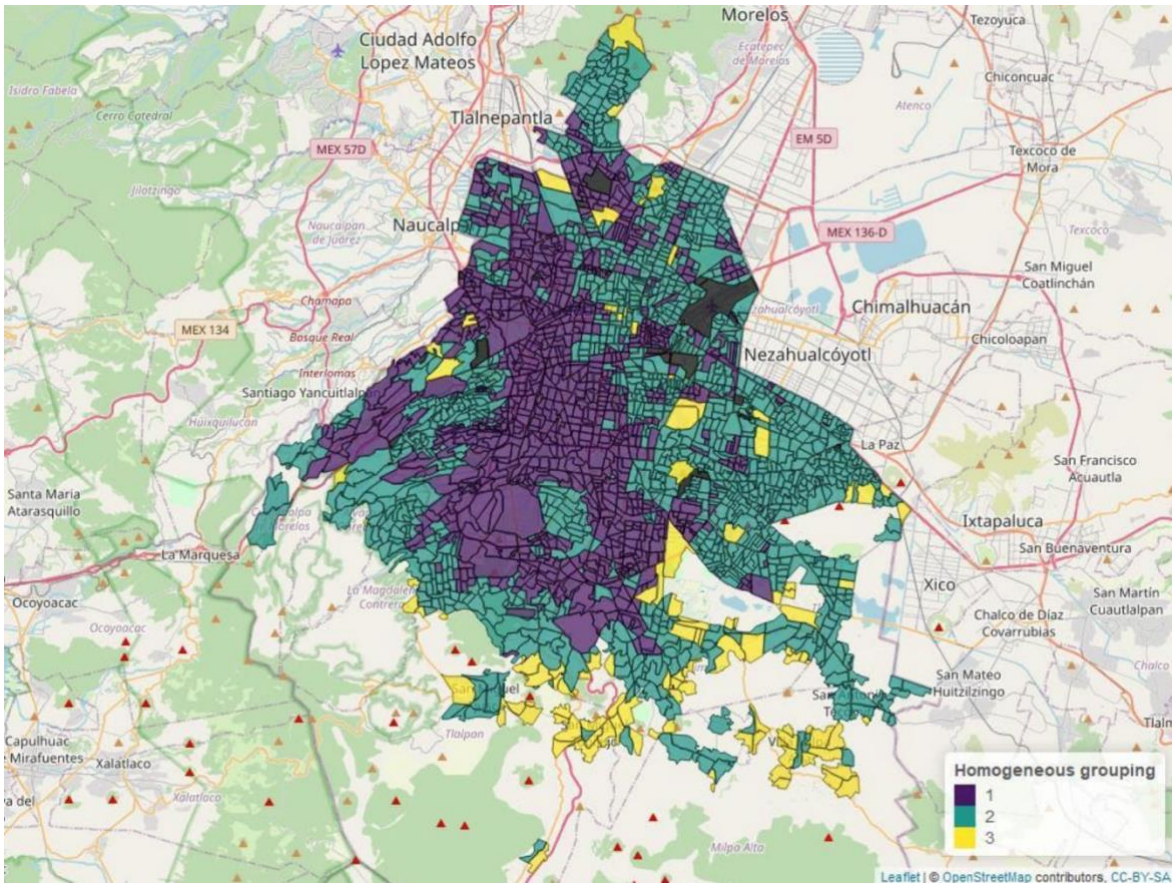

Figure 35: Homogeneous grouping of areas in Mexico City

Positive Deviant Identification

After dividing the study population into homogeneous groups comes the stage of identifying *outliers* or positive deviants within each group separately. Positive deviants are identified within homologous groups because, as mentioned earlier, it is their relative performance when compared with peers that have similar structural constraints that is important, rather than their absolute performance. This identification requires defining the techniques and cut-off points – the limits beyond which observations are considered positive deviants – which distinguish positive deviants from non-positive deviants. Depending on how performance is measured, there are several ways to identify positive deviants:

- **Univariate Analysis:** this is used in cases where only one variable (i.e. the performance measure) is used to identify outliers. This variable can be categorical i.e. pass/fail, win/lose or healthy/sick. In such cases, positive deviants are those that succeed when most fail. Alternatively, the variable can be continuous and, depending on the underlying distribution, a suitable outlier detection method can be used. For instance, if the data can be assumed to follow a normal distribution, then positive deviants can be defined as observations at the extreme end of the distribution, where the cut-off point might be two standard deviations from the mean. If no assumption of normality can be made about the underlying distribution, extreme value analysis can be used, which deals with the extreme deviations from the median of distributions (De Haan, Ferreira & Ferreira 2006). Proximity-based models can also be used (e.g. clustering and density-based methods), where outliers are points isolated from the remaining data on the basis of similarity or distance functions (Aggarwal 2013). In the Egypt research publication case study (Albanna, Handl & Heeks 2021), positive deviants in each citation metric were identified using a density-based method called the interquartile range (IQR). IQR segments an ordered dataset into quartiles and the values that separate them are denoted by Q1, Q2 and Q3 (Hampel 1974). Positive deviants were defined as observations that fall above $Q3 + 1.5*(Q3-Q1)$.

- **Multivariate Analysis:** this is used in cases where there are contextual variations among the observed units belonging to the same homogeneous group that need to be controlled for (i.e. to reduce their effect). Those contextual/structural variables are used to predict performance for each observed unit using regression analysis, and the positive deviants are identified based on how far the observed performance is from the predicted performance. This increases the likelihood that the identified positive deviants are overperforming due to individual practices and strategies and not due to structural and contextual factors that can be accounted for in the regression. When the performance measure is categorical, probabilistic models such as logistic regression can be used. Positive deviants in this case are the false negatives i.e. observations that based on the independent variables are expected to fail but in fact succeeded. When dealing with continuous performance measures a least-squares fit is typically used (Aggarwal 2013). In the Ecuador cattle-farming project, we used a model to predict farm deforestation rates as a function of farm cattle density, size, soil adaptability, socio-economic variables and the different land uses (Grijalva et al. 2021). Positively-deviant farms were then identified based on the residual values i.e. the difference between predicted and observed deforestation rates.

- **Posteriori Expectation:** a phenomenon based on historical observation is the basis on which the cut-off point is determined. For example, according to the International Union for Conservation of Nature, threatened species are defined as species that suffer a decline in population for three generations, or over 30 years. A positive deviant could be a population of a species whose size is increasing, or is stable, for three generations or more, when the size of other populations of the species is decreasing rapidly.

- **Exceptional Responders:** exceptional responders are units that perform better than expected in response to a certain intervention. An example would be an intervention to protect forests. Forest cover could be measured both inside and outside a protected area. The difference between the inside and outside can be used to generate an average expected effect of protection and positive deviants would be the protected areas significantly exceeding the expected effect. This can be done using the difference-in-

differences method (Abadie, Diamond & Hainmueller 2010), where positive deviants would be the units having the largest difference in differences.

Positive Deviant Validation

The previous step aims to identify outliers. However, one risk of using non-traditional data is that it is possible to find "spurious correlations" or, more generally, to misinterpret statistical outliers as positive deviants. This spuriousness is usually caused by confounding factors that the data could not capture, or due to making wrong comparisons in the homogeneous or peer grouping i.e. not comparing like with like (Blastland & Dilnot 2008). Hence, potential positive deviants identified in the previous stage should be approached as a starting point for asking questions rather than as the basis for drawing conclusions. Field research will be needed to ascertain if these are indeed positive deviants. However, there are ways to validate these potential positive deviants before going to the field. We generally recommend reaching out to community leaders, government officials, local domain experts and development professionals who are engaged in activities, projects or services related to the targeted areas before doing the field research. Sharing with them the initial list of potential positive deviants could lead to an early, better understanding of performance, and insight into factors that might have been overlooked or that could have biased results. For instance, there might be development interventions just for positive deviants, such as external support, which can explain their outperformance but which cannot be known from the digital dataset. Additionally, checking if significant contextual predictors of positive deviant performance (e.g. type of irrigation, month of rainfall, age demographics) are in accordance with existing literature and local domain knowledge, could count as a means of validation in itself.

There are also more quantitative ways to validate whether what is identified is simply random noise or false positives, or whether it is a sign of actual positively-deviant performance. One way is to look longitudinally (if data is available) and see if the identified positive deviants outperform over time, or whether their outperformance is a one-off event. In the Indonesia agricultural project, a time series analysis was conducted to see if the performance of rice farming villages was independent of climatic patterns over time compared to non-positively-deviant villages. This was done by developing a model to predict village average enhanced vegetation index (EVI) as a function of precipitation and temperature in 2013 by training it

using historic climate and EVI data from 2000 until 2012. The observed performance of positively-deviant villages was significantly higher than the observed performance of non-positively-deviant villages. This implies that outlier villages have likely adopted specific approaches and practices that others have not, and have established production systems that delink climatic patterns and productivity. This provided an initial validation of their positive deviance.

Other validation methods include trying out different sources of data and different techniques to identify positive deviants. Continuing on the Indonesia agricultural project, we used both univariate and multivariate outlier detection techniques to identify potential positive deviants (Albanna, Dhar Burra & Dyer 2020), and in the Ecuador cattle-farming project, we modelled deforestation rates using both yearly predictors and interannual variations in predictors. In both case studies, there was greater confidence in the validity of positive deviants that were identified across multiple approaches[40]. An alternative approach could be to use a different dataset. For example, in the Somalia rangeland project, the use of the remote sensing datasets was complemented by the use of open-source high-resolution imagery available from Google Earth for pre-fieldwork visual inspection. The latter was used to rule out false-positive deviants in the former analysis whose vegetation scores were inflated by interventions (e.g. government reserves), and to look for early signs of pastoral and agro-pastoral activities, visible soil and conservation techniques, and other rangeland management practices. It was also used to check if the area was actually occupied by permanent or semi-permanent settlements or not at all. Through this remote inspection, as illustrated in Figure 36, we identified patterns indicating the existence of soil and water conservation techniques at a number of potential positively-deviant communities (Abdullahi, Albanna & Barvels 2021).

[40]Consistency across time, analysis techniques and data sources can all indicate that the identified positive deviants are not random errors, hence, increasing the likelihood of them being true positive deviants. However, this does not totally exclude the possibility that their outperformance could be due to externalities that were not controlled for in the identification and were not accounted for in the validation.

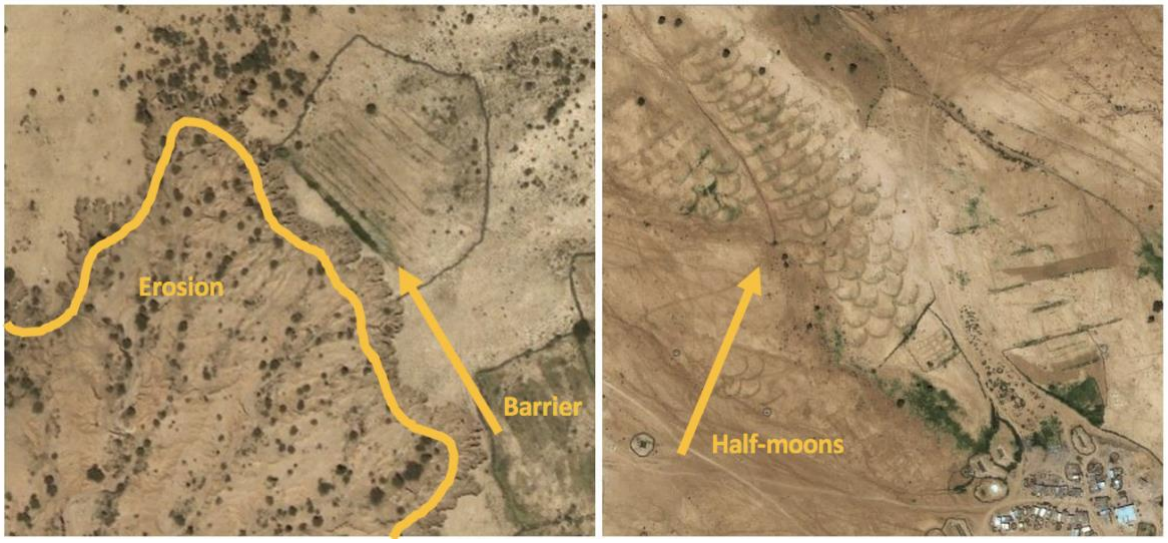

Figure 36: Examples of soil and water conservation techniques (On the left, there is a shrub barrier in the frontline with soil erosion to limit its expansion. On the right can be seen half-moon techniques to reduce water run-off. (Source: Abdullahi et al. 2021))

5.4.3 Stage 3: Discovering Predictors of Positive Deviant Performance

This section outlines the different steps needed to discover the factors underlying positive deviance. It follows the "Determining Positive Deviants" stage which results in a list of potential positively-deviant units that will be included in the fieldwork sample for further inquiry. The inquiry in this stage refers to the process of finding positive deviants' uncommon but successful strategies and practices that can be shared and acted upon by the population of interest. It starts with fieldwork planning, followed by data collection and ends with data analysis.

Fieldwork Planning

The goal of the fieldwork is twofold: 1) to confirm the validation of positive deviants identified in the previous stage, and 2) to uncover the underlying factors responsible for their deviance. The latter should include other stakeholders who have an indirect or direct relationship with the unit of analysis, and could influence its performance. For example, in the Ecuador cattle-farming project, our unit of analysis was cattle-raising farms and this stage therefore targeted both farm owners and farm workers as direct stakeholders, with government officials

identified as indirect stakeholders. Fieldwork planning should therefore start with a scoping activity: identifying the different stakeholders, and developing further familiarity with the social and cultural environment of the targeted population.

Conceptual Framework: Before developing the data collection tools, it is necessary to identify relevant variables for the field study, and to understand how they might relate to each other, and how they will be measured. One way in which this may happen is through use of existing conceptual or theoretical frameworks from the literature that have been used to explain the investigated phenomenon. Where they exist, such frameworks will likely already have been identified in Stage 2 as part of mapping relevant PD outcome and contextual variables. They may otherwise need to be created at this point. Having linked variables through a conceptual framework, this can then be discussed with key informants in the project domain and with the actors involved in the previous stage to make sure that all relevant variables are included. This particularly helps ensure that any contextual variables used in positive deviant identification that might require field validation, will be included in the data collection tools. Figure 37 presents the example of a framework that was used as the basis to develop the questionnaire tool in the Ecuador cattle-farming project.

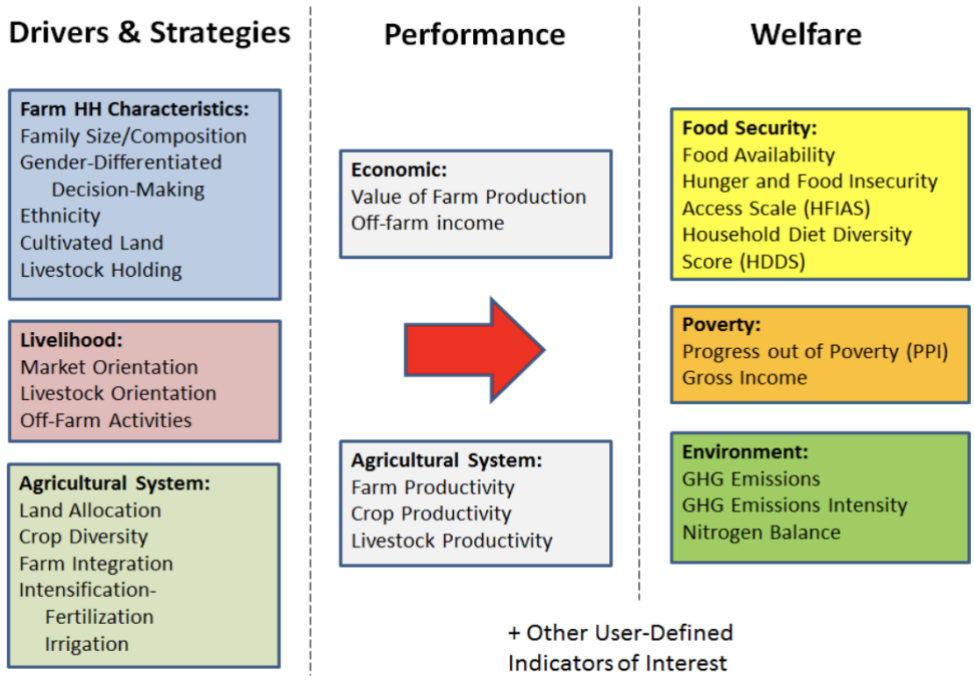

Figure 37: Conceptual framework of key farm livelihood indicators (Wijk et al. 2016)

Study design: After developing the conceptual framework and mapping out the different stakeholders, the strategy for collecting data from those stakeholders must be determined. This can use a qualitative approach (e.g. interviews), a quantitative approach (e.g. surveys) or a mix of both. In the following 'Data Collection' step we will present the different methods that can be used in each of those approaches. However, due to the nature of the DPPD method – which covers populations that are relatively larger than in the conventional PD approach – a mixed-methods approach is likely most appropriate. This is because it supports the combined analysis of a small information-rich sample of positive deviants to qualitatively generate hypotheses about individual, cultural, social and structural predictors of positive deviant performance via inductive reasoning, while also leveraging large samples to validate the generated hypotheses quantitatively via deductive reasoning. Figure 38 presents the proposed mixed methods study design for DPPD projects. This was used, for example, in the Egypt research publication case study, where positively-deviant researchers were interviewed first to generate hypotheses about the basis for their performance, and then quantitative data were collected from both positive-deviant researchers and nonpositive-deviant researchers to validate those hypotheses and identify significant differences between both groups (Albanna et al. 2021).

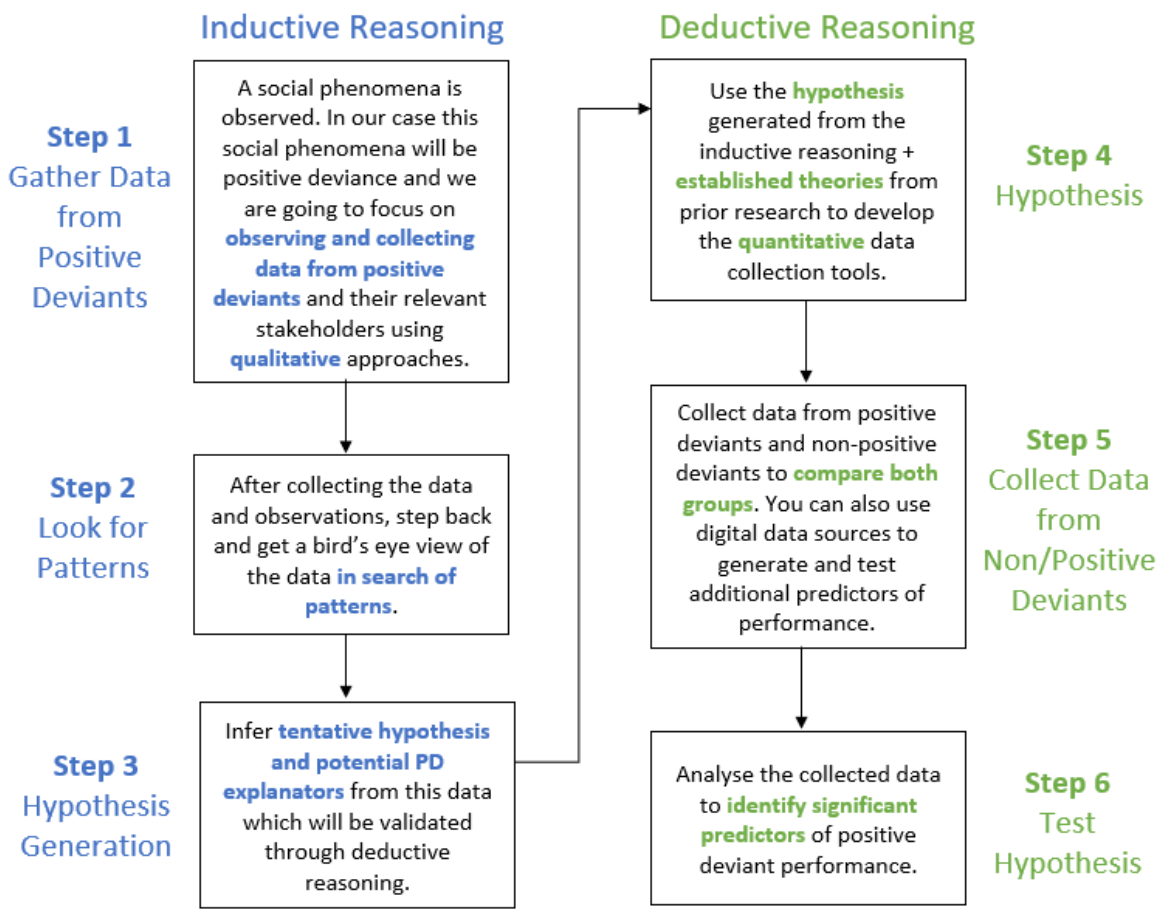

Figure 38: DPPD study design

The first step of Figure 38 could also include a few non-positive deviants to establish an understanding of the population's normative and common behaviour before interviewing positive deviants; hence, making it easier to identify uncommon practices and strategies. However, it is advised to build this normative framework at an earlier stage of the study ('Assessing Problem-Method Fit') because it might uncover key variables that are needed in determining positive deviants. Furthermore, in cases when there are limited resources constraining the ability to conduct large scale surveys, following a qualitative approach targeting both positive deviants and non-positive deviants rather than a mixed-methods approach could be more practical. Conversely, when doing retrospective studies using secondary data sources or when it is difficult to have face-to-face engagement with the study participants, a quantitative approach could be more suitable.

Data Collection

This step involves collection of the data needed to identify factors that enable positive deviants to achieve better outcomes than their peers. While the conventional PD approach focuses mainly on individual-level factors, the DPPD method – due to its breadth of coverage – can employ a 'systemic' lens that takes into account factors beyond individuals (e.g. infrastructure, policies, social system dynamics, etc.). This enables DPPD to capture a more comprehensive understanding of the complex forces at play behind a 'solution' and can lead to both community-level and policy-level interventions; though at both these and individual levels, the focus should always be on identifying factors that are transferable and controllable. Hence, it is important to design the data collection instruments in such a way as to capture this mix of factors. There are several methods that can be used for this purpose, and the choice of participants included in each method depends on the study design.

While time-consuming, *qualitative methods* provide deeper insight into the factors underlying positively-deviant performance and, as shown in Figure 38 for a mixed-methods approach, are particularly associated with generating hypotheses about positive deviants. Interviews, focus groups and observation are all relevant techniques. In the projects, we used semi-structured interviews targeting a sample of positive deviants and non-positive deviants (for example, 18 farmers – 9 positive deviants and 9 non-positive deviants – were interviewed at their farms in the Ecuador cattle-farming project). The interview schedules contained closed-ended questions and observational checklists to capture contextual, demographic and socio-economic variables in addition to variables developed from the conceptual framework. Open-ended questions were also used to uncover the uncommon strategies, attitudes and practices of positive deviants by comparing them with those of non-positive deviants.

Community-based participatory methods have been quite widely used in the conventional PD approach. They have been shown to mobilise populations, create buy-in, increase knowledge and change attitudes. In such methods the researcher acts as a facilitator who creates a space to integrate the expertise of both insiders and outsiders who contribute equally to the PD inquiry, community capacity building and action (Teufel-Shone et al. 2019).

Examples of qualitative participatory methods, the first of which has been used in the projects to date, include**:**

- **Community Mapping:** the process and product of a community getting together to map their own assets, values, beliefs, spatial units of interest or any other self-selected variable. In the Somalia rangelands projects, participatory mapping (see Figure 39) has been used to: 1) identify community resources and infrastructure; 2) understand the mobility patterns of livestock and pastoralists (i.e. transhumance); 3) understand the conditions of the rangeland and map both problem areas and bright spots of rangeland conservation; and 4) understand the different land uses and user groups in the community and areas where there is potential conflict. Similarly, feminist participatory cartography was used in the Mexico safe public spaces project, to visually represent the experiences and knowledge of women who live, work, study, visit or transit through the urban spaces that were targeted for fieldwork. 89 women across the 16 selected spaces (10 positive deviants and 6 non-positive deviants) were invited to highlight on a map where they hang out, rest, shop, among other activities, as well as places where they feel safe and places where they feel unsafe and why they feel this way about those places. Those individual area maps were then amalgamated into a collective one as shown in Figure 40.

Fig. 39 Community mapping at the village of Shilmaale

Fig. 40 Collective map of safe (green) and unsafe (red) areas for women in one part of Mexico City

- **Participatory Sketching**: a method of collective drawing employed to obtain enriched narratives from participants (Greiner, Singhal & Hurlburt 2010). Participants jointly draw a sketch describing what they envision as good practice or an ideal model in a physical space, and then share and discuss. This has been used in PD studies when visual aids are required to identify positively-deviant practices (Nieto-Sanchez et al. 2015).
- **Discovery and Action Dialogues (DADs):** a key technique used in PD, the aim of DADs is to ensure that in the presence of a facilitator, people in the group, unit, or community discover by themselves the positively-deviant practices (Escobar et al. 2017). DADs are argued to create favourable conditions for stimulating participants' creativity in spaces where they can feel safe to invent new and more effective practices; to reduce resistance to change as participants are given the freedom to choose the practices they will adopt and the problems they will tackle; and to increase the likelihood that solutions will be adopted by creating local ownership. DADs are thus seen as a basis for both discovering PD practices and mobilising communities to take action.
- **Photo Elicitation:** a method used in visual anthropology that introduces pictures to elicit comments (Lindlof & Taylor 2017). For example, pictures taken during interviews

with positive deviants can be presented to focus group participants. Using these pictures as reference, the participants are asked to reflect on the captured practices and solutions.

- **Data-Driven Participatory Approaches:** include methods that engage community members in interpreting the data that were collected about them in order to catalyse dialogue and debate around the challenges they are facing and means to address those challenges (Cañares 2020). Alongside transforming community members from passive producers of data into active users, this can be used to elicit PD-related evidence from the community.

Quantitative methods, as noted above and in Figure 6, can be used to test hypotheses about positive deviants using statistical analysis. For example, quantitative surveys can collect structured data from both positive deviants and non-positive deviants to identify statistically significant differences between both groups. Quantitative observation checklists can also be applied, containing a list of things that the observer will look at when observing positive deviants and non-positive deviants. Usually, it incorporates contextual variables that are used to identify positive deviants and require ground validation. For instance, in the Ecuadorian cattle-farming project, we used vaccination data as a proxy of cattle numbers in the farm (Grijalva et al. 2021). The field team had to validate this proxy by counting the real number of cattle. Knowing we found a good correlation between the vaccination data and actual cattle headcount, we were then able to propose that such data can likely be used for the same purpose in other studies.

The DPPD method provides an opportunity to use quantitative digital datasets not just for identification of positive deviants but also for understanding their underlying behaviours and practices. While this is not possible in most cases, it is still important to ask the question "Are there digital traces that can shed light on positive deviant behaviours and practices?". In the Egypt research publication case study, we applied machine learning and content analysis techniques to the researchers' publications to identify paper-extrinsic factors (e.g. number of pages), paper-intrinsic factors (e.g. topics covered) and publication outlets (e.g. where do they publish their research) that could shed light on publication strategies and tactics of those positively- deviant researchers whose research was highly cited (Albanna, Handl & Heeks 2021).

Data Analysis

The main aim of this step is to identify significant predictors of positive deviants that distinguish them from non-positive deviants. Data analysis techniques will largely depend on the selected study design: qualitative, quantitative or mixed-methods. At the heart of the qualitative data analysis is thematic analysis of verbatim interview and focus group transcripts to extract the attributes, attitudes, practices and strategies of positive deviants. Such analysis can also quantify the frequency of occurrence of these variables, offering some measure of difference between positive deviants and non-positive deviants. In a mixed-methods approach as per Figure 6, the themes inductively identified can be used to develop a survey instrument that seeks to quantitatively validate the qualitative findings (uncommon PD predictors) using a large representative sample of the population. For example, in the Egypt research publication case study, the qualitative analysis of the interviews with PDs led to the discovery of predictors that proved to be significant in the following quantitative analysis of the surveys targeting both positive deviants and non-positive deviants. Examples of those predictors include, but are not limited to: publishing with foreign reputable authors, and taking scientific and formal writing courses (Albanna, Handl & Heeks 2021).

PD studies use three main types of quantitative analysis: descriptive statistics, inferential statistical tests and regression analysis. Descriptive statistics are used as the first step of statistical analysis for either two-group comparison (positive deviants vs. non-positive deviants) or three-group comparison (positive deviants vs. two other groups: average performers and negative deviants who significantly underperform). Statistical tests provide basic comparative information for these groups (differences between group means, minima and maxima, etc.) and also establish whether differences are statistically significant when comparing either the two groups (e.g. via student t-test, Mann Whitney and Fisher exact test) or three groups (ANOVA, Kruskal–Wallis and chi-square). For example, in the Mexico safe public spaces project, this analysis was used to identify differences between positively-deviant AGEBs and non-positively-deviant AGEBs. Early findings revealed that positively-deviant AGEBs, in some homogeneous groups, had a higher percentage of streets with informal commerce and public lighting, more poles with cameras and more “intersecciones seguras”

interventions[41] compared to non-positively-deviant AGEBs. Regression analysis is used to examine the relationship between the identified positive deviant performance measure (as dependent variable) and the independent variables. It can include both the structural variables and controls that were used in the positive deviant identification step and the socio-demographic and behavioural variables captured in the data collection step. For each homogeneous group, we recommend having a separate model to identify significant predictors of performance that are relevant to the context of the respective groups.

5.5 Results

This section summarises the preliminary results from applying the first three stages of the DPPD method across the five pilot projects that conducted fieldwork[42]. Table 28 presents for each project the study population, the number of potential positive deviants identified from Stage 2, and potentially-transferable explanators of outperformance that were generated from the fieldwork investigation and/or the analysis of secondary data.

Project	**Study Population**	**Positive Deviants**	**Potentially-Transferable Positive Deviance Explanators**
Research publication outperformance in Egypt (Albanna, Handl & Heeks 2021)	203 information systems researchers who are affiliated to public universities	26 potential positive deviants were identified	- Taking scientific writing and English language courses - Supervising a large number of postgraduate students - Publishing with foreign authors - Establishing research teams overseas - Obtaining PhD degrees from global North universities - Securing research grants and travel funds - Publishing more journal articles and fewer conference papers - Working on established research areas rather than on radical research topics

[41] Safe intersections programme: https://www.eluniversal.com.mx/metropoli/cdmx/con-intersecciones-seguras-se-redujo-un-30-los-accidentes-viales

[42] The Indonesia agricultural project covered only the first two stages of the DPPD method.

			- Having multiple authors and affiliations in their publications
Rangeland preservation by pastoral communities in Somalia (Abdullahi, Albanna & Barvels 2021)	314 communities in the West Gollis area of Somaliland, which is a zone with a majority pastoral community	13 communities were identified as potential positive deviants	- Women are part of the village development committee and are involved in decision making - Trained community-based animal health workers - Dedicated village committees for conflict resolution - Dedicated natural resource management committees - Stone lines and stone bunds to slow down water runoff - Rotational reseeding and strip grazing of pasture mixes such as Rhodes grass and Lablab legume - Income diversification (beekeeping, power food preservation, growing fodder) - Community land enclosure policies - Customary laws for tenure management - Drip irrigation - Moving towards agro-pastoralism - Tree planting
Cereal-farming outperformance in Niger (Gluecker, Lehman & Barvels 2021)	12,093 communities in the Sahelian region, where there is a predominance of rain-fed agriculture	180 communities were identified as potential positive deviants	- Leaving millet stalks and stems in the field to protect the soil from wind erosion and help restore the organic matter - Existence of *Faidherbia albida* ("Gao tree") which helps fertilise the soil - Existence of Zai holes and stone bunds to reduce surface water run-off - Sowing only after useful rains[43] - The rational use of mineral fertiliser with organic manure
Public spaces in Mexico City where women are safer	2,431 AGEBs in Mexico City	32 AGEBs were identified as potential	- Presence of informal commerce - Better lighting (intensity, scheduling, location and distribution) - More poles with cameras - The existence of safe intersection programmes

[43] Useful rain (14 mm in fallen water height) is the point at which producers can sow usefully. Sowing only after a useful rain helps to reduce seed loss considerably.

(Cervantes, Rios & Soto 2021)		positive deviants	- Lower percentage of green areas - Presence of activities and facilities for specific demographic groups (women, children, families and elderly)
Low deforestation cattle-farming in the Ecuadorian Amazon (Grijalva *et al.* 2021)	5,332 farms in Joya de los Sachas canton and 5,701 farms in Sucúa canton	53 potential positive deviants were identified across both cantons	- Income diversification - Motivation to plant native trees - Rotational grazing - Finding other sources of animal feed which reduces pressure on pasture - Realising the value of trees in their grazing systems

Table 28 Summary of results from DPPD projects

The predictors of PD performance could be divided into those (listed in the table) that could potentially be transferred in some way, and those that could not. For example, in the Egypt research performance case study, we were able to identify potentially-transferrable practices that could enhance the publication outcomes of Southern researchers, such as participation in multi-country teams, taking specific training courses, or focusing on publishing journal articles rather than conference papers. But there were also factors contributing to better outcomes that are not transferable such as seniority, publication age or being male. Similarly, in the Ecuador cattle farming project, although we did find some attitudes and practices in favour of forest conservation that might be scaled out to other farmers, in most cases, low deforestation was a result of contextual and structural factors. For instance, in some farms there were topographic limitations where trees on ravines, swamps and mountainous zones were untouched simply due to their geographic inaccessibility and the difficulty of land use.

The potentially-transferable PD explanators emerging from the field could feed into both community and policy interventions. Explanators relevant to community interventions are mainly practices, strategies and know-how of positively-deviant individuals that could be transferred and replicated by other individuals in the group. Examples include Zai holes (holes for seeds that contain organic fertiliser and trap rain) and the knowledge of what would be "useful rain" in Niger; and rotational reseeding and strip grazing in Somalia. Explanators feeding into higher-level policy design include existing successful policy interventions,

government schemes or physical features of positively-deviant geographic units that could be adjusted to achieve the desired outcomes. For example, in the Mexico safe public spaces project, better lighting and the presence of informal commerce, especially if the merchants are known to women or are women, were associated with lower reported crime rates. These elements could inform policies on the design of urban spaces, making them safer. In the Somalia rangelands project, pastoralists stressed the importance of establishing government reserve areas. These reserves emerged as crucial for saving livestock during the dry season or droughts and, hence, their creation and effective management would be a policy recommendation.

When designing community-level interventions from these projects, we will focus on creating activities that enable people to share, learn and practice the behaviours of positive deviants. Activities will be designed to warrant the active participation of those who developed the solution i.e. positive deviants, those who stand to benefit from adopting a positively-deviant practice i.e. non-positive deviants, as well as the different stakeholders who might have an influence on the overall adoption of the solution. Examples for communities could be community gatherings where positively-deviant pastoralists demonstrate how they build the stone lines and bunds to limit soil erosion, or where they explain how village leaders limit illegal land enclosures and manage conflict in the communal areas using effective customary laws.

5.6 Discussion: Lessons Learned

Having provided details on the first three stages of the DPPD method and the results produced to date, we now draw out some of the key lessons we learnt while applying the method, as developed from reflections of both the global and country-level teams during learning calls and online surveys undertaken as part of the six projects. These reflections highlight both the limitations and opportunities of applying DPPD, and its future potential.

5.6.1 DPPD is not Universally Applicable

DPPD is not a method that could be applied to every PD-amenable problem. This is because non-traditional digital data that could be utilised in DPPD must be capable of capturing the

performance of the observed units without compromising their privacy. This is difficult to achieve in culturally-sensitive domains, such as limiting HIV transmission or fighting against female genital mutilation, where the conventional PD method has been applied successfully. Additionally, open digital data is rarely available at the level of individuals, mainly due to the prerequisite to de-identify and aggregate digital observations to make them open. This makes DPPD better suited to development problems with communities or geographical areas as the unit of analysis, with the exception of a few domains where the digital outcomes of individuals can be traced and quantified without compromising their privacy (e.g. scientific research outputs). Finally, the DPPD method relies heavily on the existence of reliable and accessible digital and secondary data that captures outcomes directly related to the addressed development problem. In domains and countries with poor data landscapes, applying the DPPD method may not yet be feasible.

5.6.2 The Right Know-How Must be Available

Finding potential positive deviants from non-traditional data without sufficiently understanding their contextual realities will likely lead to false positives. Hence, a unique combination of local, domain and data knowledge is needed before conducting any data analysis. Country-specific domain knowledge is crucial in understanding the normative behaviours of the investigated population, if positive deviants exist, and the contextual and structural factors that have an effect on their outcomes. Domain-specific data knowledge is required to identify relevant performance indicators from the available data, in addition to mapping out suitable data sources that could be used. However, such expertise is usually missing within international organisations. Therefore, an initial mapping of existing and missing relevant know-how for the project can help uncover necessities for bringing in additional know-how.

5.6.3 Control for Contextual Variables

The conventional PD method generally covers small sample sizes, e.g. a few dozen families, in a homogeneous context, e.g. a single village or district. This makes it very accurate in singling out a particular behaviour that explains a successful practice since non-behavioural factors can be largely neglected as they are more or less the same for the entire (small)

population being investigated. Digital performance measures used in DPPD can cover large geographic areas enabling the inclusion of larger populations in the analysis. This increases the heterogeneity of the sample and the likelihood of potential confounding factors when identifying positive deviants. For example, structural factors such as access to roads and levels of rainfall, and socio-economic factors such as population density, differ across large populations and could contribute to differences in performance among units of analysis. Failure to control for those structural factors when identifying positive deviants leads to an inability to single out the particular attributes, practices and strategies that need to be disseminated. The biggest challenge here is identifying additional data sources, both traditional and non-traditional, that can link the context to the digital performance measure. Additionally, such contextual data should have an overlapping time frame and spatial resolution with the performance measure to be useful. In Ecuador, we used satellite imagery to calculate deforestation rates for a large sample of cattle-raising farms. However, cattle density is an important confounding factor, as higher density makes the recovery of pasture harder, and requires more grazing space which creates pressure to deforest. We used cattle vaccination data as a proxy of cattle numbers on the farm to identify positive deviants with low deforestation rates relative to their cattle density and not in absolute terms. This increased the chances of attributing low deforestation rates to sustainable cattle-ranching practices and not to lower cattle density.

5.6.4 If Possible, Measure Performance Over Time

An advantage of using digital measures of performance is that they often have a longitudinal coverage and are collected at regular intervals. This allows performance to be evaluated over time, and to observe moves towards or away from positive deviance. Furthermore, it enables identification of persistent positive deviants: those who appear as positive deviants in the data for several consecutive years are more likely to be "true" positive deviants. In the Ecuador cattle-farming project, we were able to measure deforestation rates over a five year period. We were able to develop a more nuanced understanding of positive deviants: those who became positive deviants over time (from low performing to high performing) or those who stopped being positive deviants (from high performing to low performing). Such diversity in positive deviant categories can help uncover interesting factors that trigger moving from one

state to another and can inform the design of interventions. Moreover, the same digital datasets that are used to capture performance longitudinally can readily be used to monitor and evaluate the mid-to-long-term effects of scaling the practices and strategies of positive deviants across intervention populations.

5.6.5 Adopt a Holistic Approach in Understanding PD

The potentially-wide spatial coverage of DPPD, when compared with the conventional PD method, provides an opportunity to observe units of analysis that are beyond individuals e.g. villages or regions. Discovering determinants of outperformance within such units requires a new type of inquiry that looks at factors beyond individuals that could be modified and transferred. Such factors include, but are not limited to, governance mechanisms, development interventions, policies, systemic changes, etc. Early findings from our pilots suggest that the DPPD method might be a promising way to better understand the interactions between individual and supra-individual factors. This can inform the design of nuanced interventions that take into account such interactions, hence, increasing their effectiveness and contextual fit. This is different from the conventional PD method which is placed in a more 'controlled' environment where variation in performance might indeed be attributed only to individual-level factors. As a case in point, in the Somalia rangelands project, we realised through conversations with local experts that rangelands health is influenced by individual and community behaviours e.g. soil and water conservation techniques, alternative livelihoods, along with land tenure policies and campaigns against private enclosure. Hence, when planning for our field investigation of positively-deviant communities we decided to embrace the complex dimensions of the rangeland problem and explore positive deviance as a system behaviour instead of looking into positive deviants as individuals in isolation from the larger system. Findings from this investigation could thus inform the design of both community-level and policy-level interventions.

5.6.6 Earth Observation Data is a Low Hanging Fruit for DPPD

After applying the DPPD method to multiple projects and domains, it is clear that earth observation (EO) data can play an instrumental role in the viability and scalability of the DPPD method. EO gathers data about the physical, chemical and biological systems of the

planet using remote sensing technologies (Rast & Painter 2019). It is considered the most cost-effective technology able to provide data at a global scale. It can be acquired at low cost, over long periods of time, and thanks to the recent advances in remote sensing technologies, it is witnessing a growing availability at a high resolution including coverage of lowest-income countries where other datasets are lacking. Such attributes of EO data make it possible to overcome a number of data accessibility limitations, while being able to capture the potential gains of using big data in PD such as reducing the cost, time and risk of measuring performance at large scale. Of course, limitations must be acknowledged given that EO data is applicable only for problems where the impact of human behaviours and practices on natural and built environments can be observed and measured remotely. For instance, EO data has proved useful in our projects to identify positive deviance in vegetation health and forest cover (assuming that this observed deviance can be linked to individual practices and strategies, or successful policies and governance mechanisms on the ground). Additionally EO data was useful in the homogeneous grouping step, where remote sensing-derived covariates (e.g. temperature) were extracted to create peer groups having similar conditions. High resolution satellite imagery also proved useful in validating potential positive deviants by ensuring the accuracy of the land covers used for PD identification and in identifying hints of positively-deviant practices in the rare case that they are observable.

5.7 Conclusion

This paper has presented the three core stages of the data-powered positive deviance (DPPD) method; a new way of applying the positive deviance approach by combining non-traditional, digital data (e.g. online and remote sensing data) with traditional data (e.g. interviews, official statistics). These core stages are: assessing problem-method fit, determining positive deviants, and discovering positive deviant practices and strategies. The remaining two stages covering the design of interventions and monitoring and evaluating the effects of those interventions were not included in the presented method for two reasons: the majority of the projects reported here did not yet reach these stages, and these stages should not differ much from the conventional PD approach. However, investigation of the potential value-added benefits that could be incorporated into those two stages from the use of non-traditional data is a future direction of this work.

More generally, the DPPD method makes it possible to identify and characterise positive deviants at temporal and geographical scales that are not possible using the conventional approach. While the use of existing datasets may reduce initial time/financial costs of PD identification compared with traditional PD methods, DPPD overall is not yet demonstrably cheaper and quicker because there can be additional costs associated with the access and analysis of datasets, because DPPD itself will typically involve fieldwork, and because none of the pilot projects is yet in a position to allow total and comparative lifecycle costs to be calculated. The presented method was developed iteratively through its application by the DPPD initiative partners in six projects across five different development domains. Through readily available digital data we were able to observe and capture outcomes of large populations in relation to the addressed development problems; however, this came with the challenge of controlling for numerous contextual factors, parts of which were feasible while others not. The large temporal coverage of digital data enabled not only the identification of sustained positively-deviant behaviour (e.g. in consecutive years) but also changes in behaviours (i.e. becoming positive deviants or no longer being positive deviants). Additionally, accessing relevant data turned out to be much harder than expected. This highlights the necessity to forge the right partnerships and involve the various data-controlling stakeholders at an early stage of the project.

The DPPD method relies heavily on a digitally recorded or observed performance measure that is directly related to the desired outcomes of the observed units. However, the selection of this measure highly depends on the data availability in a given country and domain. Flexibility and creativity in dealing with a potential lack of data, while adhering to the original focus of the development challenge, requires constant iteration, reflection and discussion with domain experts. It is also evident that DPPD requires an interdisciplinary team which combines the right local, domain and data analysis know-how to conduct plausible data analysis for PD identification that uses relevant performance measures while controlling for potential confounding factors. That specialist expertise can be in short supply and one future aim would be to 'democratise' the method, enabling it to be accessible by a wider range of organisations including, potentially, community-based organisations and others involved with citizen science. Finally, while the conventional PD approach focuses mainly on

individual-level factors, the DPPD method – due to its large spatial coverage – could employ a more holistic lens that takes into account both individual and supra-individual factors.

We hope that the details of the DPPD method provided here enable its uptake by development and data science professionals, and we encourage its application to a wider range of development challenges and in a wider set of development domains, with further refinement of both the method and the lessons learned.

Acknowledgements

We would like to thank Catherine Vogel from the GIZ Data Lab for leading the effort to establish the DPPD initiative. We thank the in-country United Nations Development Programme Accelerator Lab teams, the Deutsche Gesellschaft für Internationale Zusammenarbeit projects and all their local partners for implementing the Ecuador, Mexico, Niger and Somalia projects. We would specifically like to thank Ana Grijalva, Paulina Jimenez and Carolina Michell from the Ecuador project; Gabriela Rios, Alejandra Cervantes and Daniella Torres from the Mexico project; Assane Boukar, Moustapha Sahiro, Moctar Seydou, Rachida Mani and Damien Hauswirth from the Niger project; and Hodan Abdullahi, Erik Fritzche and Florian Fritzche from the Somalia project. We would also like to thank Gunnar Hesch and Esther Barvels from the GIZ, and Erik Lehmann from the GIZ Data Lab for their support in the geographic information system and remote sensing analysis conducted in the Somalia and Niger projects. We are also grateful to the United Nations Global Pulse Lab Jakarta and their local partners for implementing the Indonesia project. This work was financially supported by the GIZ Data Lab and the in-country UNDP Accelerator Labs and GIZ projects.

References

Abadie, A., Diamond, A. & Hainmueller, A. J. (2010) Synthetic control methods for comparative case studies: Estimating the effect of California's Tobacco control program, *Journal of the American Statistical Association*.

Abdi, H. & Williams, L. J. (2010) Principal component analysis, *Wiley interdisciplinary reviews: computational statistics*. Wiley Online Library, 2(4), 433–459.

Abdullahi, H., Albanna, B. & Barvels, E. (2021) *Rangelands Defying the Odds: A Data Powered Positive Deviance Inquiry in Somalia*, *Data Powered Positive Deviance*. Available at: https://dppd.medium.com/rangelands-defying-the-odds-a-data-powered-positive-deviance-inquiry-in-somalia-90772de392dd (Accessed: 4 August 2021).

Aggarwal, C. C. (2013) *Outlier Analysis*. New York, NY: Springer New York.

Albanna, B., Dhar Burra, D. & Dyer, M. (2020) *Identifying Potential Positive Deviants (PDs) Across Rice Producing Areas in Indonesia: An Application of Big Data Analytics and Approaches*. Jakarta: Pulse Lab Jakarta.

Albanna, B., Handl, J. & Heeks, R. (2021) Publication outperformance among global South researchers: An analysis of individual-level and publication-level predictors of positive deviance, *Scientometrics*, 1-57.

Albanna, B. & Heeks, R. (2019) Positive deviance, big data, and development: A systematic literature review, *The Electronic Journal of Information Systems in Developing Countries*, 85(1), e12063.

Ang, Y. Y. (2019) *Integrating big data and thick data to transform public services delivery*. Washington DC: IBM Center for The Business of Government.

Blastland, M. and Dilnot, A.W., (2008) *The Tiger that Isn't: Seeing Through a World of Numbers.* London, UK: Profile Books.

Bornakke, T. & Due, B. L. (2018) Big–Thick Blending: A method for mixing analytical insights from big and thick data sources, *Big Data & Society*, 5(1), 205395171876502.

Bradley, E.H., Curry, L.A., Ramanadhan, S., Rowe, L., Nembhard, I.M. and Krumholz, H.M., (2009) Research in action: using positive deviance to improve quality of health care, *Implementation Science*, 4(1), 1-11.

Cañares, M. (2020) *Three examples of data empowerment*, *Data Empowerment*. Available at: https://medium.com/data-empowerment/three-examples-of-data-empowerment-5f3e964ffbdc (Accessed: 21 August 2021).

Cervantes, A., Rios, G. & Soto, I. (2021) *Identifying Safe(r) Public Spaces for Women in Mexico City*, *Data Powered Positive Deviance*. Available at: https://dppd.medium.com/identifying-safe-r-public-spaces-for-women-in-mexico-city-4f3d49d269d6 (Accessed: 4 August 2021).

Data Powered Positive Deviance (2020) *Launching the Data Powered Positive Deviance Initiative*. Available at: https://dppd.medium.com/.

De Haan, L., Ferreira, A., Ferreira, A., (2006) *Extreme Value Theory: An Introduction*. New York, NY: Springer.

Escobar, N.O., Márquez, I.V., Quiroga, J.A., Trujillo, T.G., Gonzalez, F., Aguilar, M.G. & Escobar-Perez, J., (2017) Using positive deviance in the prevention and control of MRSA infections in a Colombian hospital: a time-series analysis, *Epidemiology and Infection*, 145(5), 981–989.

Gluecker, A., Lehman, E. & Barvels, E. (2021) *Searching for Positive Deviants Among Cultivators of Rainfed Crops in Niger*, *Data Powered Positive Deviance*. Available at: https://dppd.medium.com/searching-for-positive-deviants-among-cultivators-of-rainfed-crops-in-niger-8dbbcceaf4ec (Accessed: 4 August 2021).

Greiner, K., Singhal, A. & Hurlburt, S. (2010) 'with an antenna we can stop the practice of female genital cutting': a participatory assessment of ASHREAT AL AMAL, an entertainment-education radio soap opera in sudan, *Investigación & Desarrollo*, 15(2).

Grijalva, A., Jiménez, P., Albanna, B. & Boy, J. (2021) *Deforestation, Cows, and Data: Data Powered Positive Deviance Pilot in Ecuador's Amazon*, *Data Powered Positive Deviance*. Available at: https://dppd.medium.com/deforestation-cows-and-data-data-powered-positive-deviance-pilot-in-ecuador-s-amazon-648aa0de121c (Accessed: 3 August 2021).

Hampel, F. R. (1974) The influence curve and its role in robust estimation, *Journal of the American Statistical Association*, 69(346), 383–393.

Hardin, G. (1968). The tragedy of the commons: the population problem has no technical solution; it requires a fundamental extension in morality. *Science*, *162*(3859), 1243-1248

Hauswirth, D., Yaye, H., Soumalia, A. S., Djariri, B., Lona, I. & Abba, M. B. (2020) *Support for the Concerted Formulation of SPN2A for the Republic of Niger: Identification and Assessment of Climate-Smart Agriculture Options Priority for Adaptation to Changes Climate in Niger (Volume 1)*. Niamey, Niger. Baastel - BRL - ONFI. Brussels, Belgium.

Hilbert, M. (2016) Big data for development: a review of promises and challenges, *Development Policy Review*, 34(1), 135–174.

Kovács, J., Kovács, S., Magyar, N., Tanos, P., Hatvani, I. G. & Anda, A. (2014) Classification into homogeneous groups using combined cluster and discriminant analysis, *Environmental Modelling and Software*.

Lindlof, T. R. & Taylor, B. C. (2017) *Qualitative communication research methods*. Sage publications.

Mankiw, N. G. (2012). *Principles of Microeconomics* (6th ed.). Mason, OH: South-Western Cengage Learning.

Nathan, R. J. & McMahon, T. A. (1990) Identification of homogeneous regions for the purposes of regionalisation, *Journal of Hydrology*.

Nieto-Sanchez, C., Baus, E. G., Guerrero, D. & Grijalva, M. J. (2015) Positive deviance study to inform a chagas disease control program in southern Ecuador, *Memorias do Instituto Oswaldo Cruz*, 110(3), 299–309.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge, UK: Cambridge University Press.

Pascale, R., Sternin, J. & Sternin, M. (2010) *The Power of Positive Deviance: How Unlikely Innovators Solve the World's Toughest Problems*. Boston, Massachusetts: Harvard Business Press.

Positive Deviance Initiative (2010) Basic field guide to the positive deviance approach, *Tufts University*.

Post, G. & Geldmann, J. (2018) Exceptional responders in conservation, *Conservation Biology*, 32(3), 576–583.

Rast, M. & Painter, T. H. (2019) Earth observation imaging spectroscopy for terrestrial systems: An overview of its history, techniques, and applications of its missions, *Surveys in Geophysics*. Springer, 40(3), 303–331.

Sethi, V., Sternin, M., Sharma, D., Bhanot, A. & Mebrahtu, S. (2017) Applying positive deviance for improving compliance to adolescent anemia control program in tribal communities of India, *Food and Nutrition Bulletin*, 38(3), 447–452.

Smets, A. & Lievens, B. (2018) Human Sensemaking in the Smart City: A Research Approach Merging Big and Thick Data, *Ethnographic Praxis in Industry Conference Proceedings*, 2018(1), 179–194.

Somekh, B. (1995) The Contribution of Action Research to Development in Social Endeavours: a position paper on action research methodology, *British Educational Research Journal*.

Sternin, J. (2002) Positive deviance: a new paradigm for addressing today's problems today, *The Journal of Corporate Citizenship*, 57–63.

Sternin, J. & Choo, R. (2000) The power of positive deviancy, *Harvard Business Review*, 78(1), 14-15.

Sternin, M., Sternin, J. & Marsh, D. (1998) *Designing a Community-Based Nutrition Program Using the Hearth Model and the Positive Deviance Approach: A Field Guide.* Westport, CT: Save the Children.

Sternin, M., Sternin, J. & Marsh, D. L. (1997) Rapid sustained childhood malnutrition alleviation through a positive-deviance approach in rural Vietnam: preliminary findings in *The hearth nutrition model: Applications in Haiti, Vietnam, and Bangladesh Wollinka. OKeeley E, Burkhalter RB, Bashir N, (eds.). Report of a technical meeting at World Relief Corporation Headquarters*, 49-61.

Taylor, L. & Broeders, D. (2015) In the name of development: Power, profit and the datafication of the global South, *Geoforum*, 64, 229–237.

Teufel-Shone, N.I., Schwartz, A.L., Hardy, L.J., De Heer, H.D., Williamson, H.J., Dunn, D.J., Polingyumptewa, K. and Chief, C. (2019) Supporting new community-based participatory research partnerships, *International Journal of Environmental Research and Public Health*, 16(1), 44.

Trivedi, S., Pardos, Z. A. & Heffernan, N. T. (2011) Clustering students to generate an ensemble to improve standard test score predictions. In *International onference on artificial intelligence in education* (pp. 377-384). Springer, Berlin, Heidelberg.

Trivedi, S., Pardos, Z. A. & Heffernan, N. T. (2015) The utility of clustering in prediction tasks, *arXiv preprint arXiv:1509.06163*.

Vilalta, C. & Muggah, R. (2016) What explains criminal violence in Mexico City? A test of two theories of crime, *Stability: International Journal of Security and Development*, 5(1).

Wijk, M.T.V., Hammond, J., Etten, J.V., Pagella, T., Ritzema, R.S., Teufel, N. & Rosenstock, T.S. (2016) *The rural household multi-indicator survey (RHoMIS): A rapid, cost-effective and flexible tool for farm household characterisation, targeting interventions and monitoring progress towards climate-smart agriculture.* The Netherlands : CCAFS

Wallace, M.E. and Harville, E.W., (2012) Predictors of healthy birth outcome in adolescents: a positive deviance approach, *Journal of pediatric and adolescent gynecology*, 25(5), 314-321.

Wijk, M.T.V., Hammond, J., Etten, J.V., Pagella, T., Ritzema, R.S., Teufel, N. and Rosenstock, T.S., (2016) The Rural Household Multi-Indicator Survey (RHoMIS): A rapid, cost-effective and flexible tool for farm household characterisation, targeting interventions and monitoring progress towards climate-smart agriculture. CCAFS Info Note.

Wishik, S. M. & Van Der Vynckt, S. (1976) The use of nutritional 'positive deviants' to identify approaches for modification of dietary practices, *American Journal of Public Health*, 66(1), 38–42.

Zeitlin, M. (1991) Nutritional resilience in a hostile environment: positive deviance in child nutrition, *Nutrition Reviews*, 49(9), 259–268.

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P., Goodman, A., Hollander, R., Koenig, B.A., Metcalf, J. & Narayanan, A. (2017) Ten simple rules for responsible big data research, *PLOS Computational Biology*, 13(3), e1005399.

Chapter Six: Discussion

The Data-Powered Positive Deviance (DPPD) method presented in this study provides a new way of identifying and understanding positive deviance by combining big data with traditional data. It is not only an innovation in the application of the positive deviance approach, but also contributes to the big-data-for-development body of knowledge by providing a systematic way to mix analytical insights from big and thick data to identify and scale locally sourced solutions to development challenges.

The research questions used to structure this study were: (a) How can we use big data in the positive deviance approach? (b) What value will result from the use of big data in the positive deviance (PD) approach? (c) What are the challenges of using big data in the PD approach? In this chapter I draw together the findings, with specific reference to each of the research questions. Finally, I set out the contributions of this study and make recommendations for policy, practice and future research direction.

6.1 How Can We Use Big Data in the Positive Deviance Approach?

The aim of this study is to develop a method that guides the use of big data in the positive deviance approach. This method was developed iteratively through its application in one case study and five action research projects covering five different development domains. As discussed in section 6.3, big data cannot be applied to every aspect of positive deviance but, in this section, I will summarize the conditions and techniques that might ensure the successful use of big data in the three core stages of the DPPD method as outlined in Chapter Five: 1) Assessing problem-method fit, 2) Determining positive deviants and 3) Discovering predictors of PD performance.

Assessing problem-method fit: The PD approach starts by defining the problem and making sure it is an adaptive problem requiring a behavioural change. The next step would be to identify the normative behaviours of the studied population, what will be considered a desired outcome, and if positive deviants exist. In DPPD it is crucial to take this a step further and validate if the use of big data and other sources of non-traditional data could be part of the solution. In other words, can we use big data and other digital data sources to accurately

capture the outcomes of the observed units? This validation requires a unique set of local, domain and data know-how. Country-specific domain knowledge is needed to identify all the confounding structural and contextual factors that can have an effect on the outcomes captured by the digital performance measure. Those factors should be controlled for, to ensure that positive deviants are outperforming regardless of these factors' existence and not because of it. Domain-specific data knowledge is then needed to map out two types of data that will be included in the analysis: 1) a digitally observed or recorded performance measure, and 2) contextual data that can be used as controls. Those two data types should also have an overlapping temporal and geographical coverage to be deemed useful, while also having a spatial resolution that could enable linking the performance of the units of analysis with their contextual realities. Ensuring that this data is available and accessible is extremely important before starting any data analysis. For example, in the Indonesia agricultural project presented in Chapter Four, before conducting the analysis we needed to make sure that we have access to data on the household agricultural production systems, demography and socio-economic conditions. Based on this data, we decided to use the village as our unit of analysis because this was the lowest level of aggregation where it was possible to integrate the contextual data with the remote sensing-derived performance measure.

Determining positive deviants: This stage constitutes four main steps as outlined in Chapter Five: 1) performance measurement, 2) homogenous grouping, 3) PD identification, and 4) PD validation. For performance measurement, conventionally, positive deviants are identified using an outcome measure that is manually collected. DPPD suggests replacing this manually recorded measure with a digitally recorded and observed measure extracted from non-traditional data sources, such as big data, to reduce the initial time and cost needed to identify positive deviants. For instance, in the Egypt research publication case study presented in Chapter Three, digitally and readily available citation metrics were used to identify outperforming Egyptian information systems researchers. And in the Indonesia agricultural case study, openly available remote sensing-derived vegetation indices were used as a proxy of yields to identify outperforming rice-farming villages. This use of big data-derived measures of performance comes with a number of risks and uncertainties that should be minimized to ensure that true positive deviants have been identified i.e. that they are performing well due to uncommon behaviours and strategies and not due to a measurement

error. Among those risks is that such digital measures are usually proxy measures that indirectly capture the outcomes of the observed populations assuming a linear-relationship between the proxy measure and the direct measure exists, which is not necessarily the case. Hence, proxies reinforce the reliance on technical domain knowledge to assess the linearity of the relationship between the proxy indicator and its endpoint, and the reliance on ground truthing to validate this linearity, hence ensuring the reliability of using it as an alternative measure of performance. It is also possible to rely on existing literature that proves that this linearity exists, such as the case of remote sensing vegetation indices as a valid proxy of yields (Tucker and Sellers 1986; Mkhabela et al. 2011; Bolton and Friedl 2013; Johnson 2014). Additionally, in most of the cases, big data-derived performance measures cover large populations, which entails more heterogeneity and more potential confounds of performance affecting the accuracy of the identified PDs. To minimize this heterogeneity, *homogenous grouping* is extremely important to divide the study population into clusters having a similar context within which positive deviants are identified. An example is the homogenous grouping conducted prior to PD identification in the Indonesia agricultural project, where villages were divided into clusters having similar biophysical conditions as shown in Chapter Four. By doing so, positive deviants are identified in a relative sense and not in an absolute sense. Following the homogenous grouping, comes the *PD identification*. DPPD advocates the identification of positive deviants using regression models that predict performance based on a set of contextual and structural input variables. The difference between what is predicted and what is observed (i.e. the residuum) is then calculated to be the basis for the data distribution from which positive deviants are identified. An example of such multivariate analysis is presented in the Indonesia agricultural project, where potential positive deviants were identified based on the difference between what was predicted using the partial least square regression and what was observed using the EVI remote sensing-derived measure. The predictor variables were village contextual factors extracted from the agricultural census and the village potential survey. Positive deviants were defined as observations that lie the furthest (in the positive direction) from what is predicted. Big data also comes with large amounts of irrelevant data (called noise), and therefore it is necessary to conduct a *PD validation* step that filters noise from signals following the PD identification. This in particular was central to this study and will be pertinent for future applications of DPPD. The validation of the identified positive deviants was done mainly in two ways as previously

presented in Chapter Five: 1) stakeholder validation where individuals from the ground verify if the identified potential positive deviants should be targeted for the qualitative inquiry; and 2) data-based validation where we triangulate information from several data sources to confirm that those identified are true positive deviants; for example, as with the use of high resolution imagery to check for signs of human intervention in PD communities or checking if PDs persistently appear using different predictive models.

Discovering predictors of PD performance: In this stage, uncommon attributes, attitudes, practices and strategies of positive deviants are uncovered through mainly through primary fieldwork, though with some potential for big data to contribute. The list of potential positive deviants identified in the previous stage form the basis for the fieldwork sampling, where they are mainly targeted using qualitative data approaches to generate hypotheses about their behaviours. Those hypotheses are then validated using quantitative approaches that target a larger sample containing both positive deviants and non-positive deviants to make sure that the predictors of positive deviance are significant. For example, in the Egypt research publication case study presented in Chapter Three, the positively deviant researchers were first interviewed to generate hypotheses about their outperformance. Those hypotheses were then used to inform the design of the questionnaire tool that targeted a larger sample of positive deviants and non-positive deviant researchers to identify significant differences between both groups. This stage is traditionally referred to as the positive deviance inquiry where the community, having identified positive deviants, sets out to identify the behaviours, strategies and attributes that make positive deviants who they are. While this is possible in the conventional approach due to its small sample size, this could be problematic in DPPD since big data-derived measures cover a large population. This means that a larger sample of positive deviants - across a large geographic range - is identified, making the community-based participatory methods harder. This is why more traditional data collection methods were employed to identify significant predictors of PD performance, such as surveys, focus group discussions and interviews. Additionally, the nature of the DPPD method, which leverages big data in the PD approach, provides an opportunity to use quantitative digital datasets not just for identification of positive deviants but also for understanding their underlying behaviours and practices. There are cases where we used non-traditional data

sources to understand positive deviants, as when we used the publications of the researchers in the Egypt case study and open spatial databases in the Mexico safe public spaces project.

Big data sources turned out to be of greater value at the early stages of the PD approach, namely the PD identification, and of less value at the later stages of the PD approach such as the PD inquiry, where thick data collected through qualitative and ethnographic methods is better able to explain the why and how of what we are able to observe using big data. In the early stages DPPD makes use of the large scale data to observe and filter outperforming units (positive deviants), reducing (in the later stages) the qualitative search space which requires extensive time and effort to uncover the factors driving the PD performance. Figure 41 shows how the reliance on non-traditional data reduces as we go further down in the DPPD stages, similar to the population size, which gets smaller as we move further in the DPPD method.

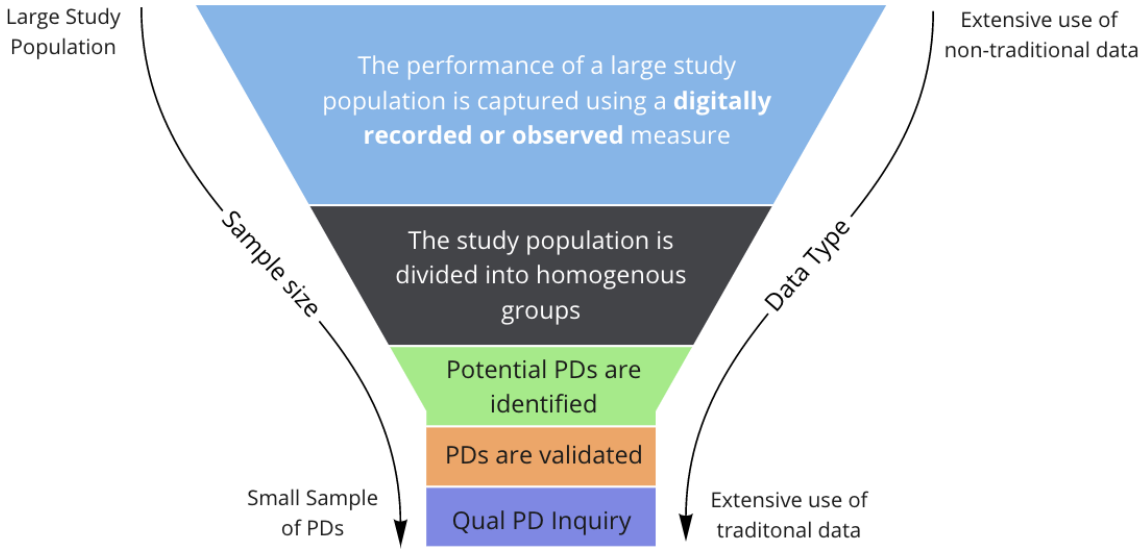

Figure 41: The DPPD funnel

The last two stages of the PD approach concern the design of interventions to disseminate PD practices and strategies, and the monitoring and evaluation of the effects of those intervention. These fell outside the scope of this study because they would have required substantial time and resources that are beyond the capacity of this PhD. In any case, stage four would not differ much from how it is applied in the conventional PD approach, but one implication from the use of big data would be to make sure that the scaling and amplification of PD practices and strategies is being done for each homogenous group separately. For

instance, in the Ecuador cattle-farming project, the study population was divided into two regions, Joya de los Sachas (north county) and Sucúa (south county), which had very different landscapes. When identifying the PD practices and strategies we found out that the sustainable practices that are relevant for Sucúa might not be that relevant for Joya de los Sachas, hence when scaling those practices the homologues that they emerged from should be considered. For stage five, potential benefits of using big data have been demonstrated, even if not yet applied; for example, the same measure that was used to capture relevant performance indicators could relatively easily be used for monitoring and evaluating interventions.

6.2 What value will result from the use of big data in the PD approach?

The conventional PD approach generally suffers from low time-, effort- and cost-efficiency, as indicated in Chapter Two. This is because it requires intensive primary data collection and observational fieldwork, which is difficult, expensive and sometimes dangerous to conduct in certain regions of the world—especially when it is entirely unguided. This study suggests that, using secondary digital and big data sources, we can identify groups of potential positive deviants before going to the field, hence reducing these initial costs and risks. Additionally, big data sources could provide insights that (in some cases) could help us better understand and characterize positive deviants. In this section I outline some of the benefits that we were able to identify and validate from the application of DPPD in the different projects presented in Chapters Three, Four and Five.

- **Reducing time and cost of PD identification:** DPPD relies on readily available digital data that can capture the outcomes of individuals, communities and geographical spaces to identify potential positive deviants. This reduces drastically the reliance on primary data in PD identification, as they are identified mainly using secondary data. Such data also permits an early, cost-efficient examination of the existence of PDs which is a necessary prerequisite to conducting any PD project. In the Indonesia project, for example, we were able to do a preliminary detection of positive deviants one week after extracting the remote sensing-derived performance indicator, which also took one week.

So, it took us only two weeks since the beginning of the project to detect potential PD rice growing villages across Indonesia. This would take months to achieve if we were to collect data on yields of rice producing villages, especially as this data was not collected in the agricultural census we used in the Indonesia project.

- **Large geographical coverage:** Readily available digital data that can capture the outcomes of individuals, communities and spaces, usually covers large geographical areas. For instance, in most of the projects, remote sensing-derived indices were used to measure the performance of different types of vegetation. Such indices are available at a global scale making it possible to measure performance at the country level. For example, DPPD covered more than 17,000 villages in the Indonesia project, which led to the identification of a large pool of potential positive deviants (around 500 villages) which could make findings more generalizable in the later stages where predictors of PD performance are discovered. This example also demonstrates the relative rarity of positive deviants which typically falls between 0 to 10% (Marsh et al. 2004): here, the sample of positive deviants constitutes just under 3% of the population. If this case study had used the traditional PD approach on a sample of 200 villages, we would have hardly found six PDs, and this would have affected our ability to make inferences about their outperformance.
- **Longitudinal coverage:** As with geographical coverage, digital data sets capturing the outcomes of the observed units in all of the DPPD projects covered long time frames which provided a dynamic view of performance. This has been of value in three main ways: 1) ***having a more nuanced understanding of performance***: performance indicators covering long timeframes made it possible to detect performance shifts i.e. moving from and to the positive deviance state over time. In the Ecuador project, we had "becoming PD" and "no longer PD" categories. Such diversity in PD categories could shed light on factors that hinder positive deviance or factors that trigger the transition from one PD state to another; 2) ***ensuring outperformance***: positive deviants who demonstrate outperformance in consecutive years have a greater likelihood of "truly" being positive deviants. This is why in multiple projects, positive deviants were defined as units that performed exceptionally well in the last three years such as the Ecuador and Niger projects, or through a trend analysis such as the Somalia project; and 3) ***enabling the use of performance indicators having a temporal dimension***. In the Egyptian

researchers case study, positive deviants were identified using indicators that had a temporal dimension such as the *hc-index,* the *aw-index* and the *m-quotient*. This would not have been possible but for the longitudinal coverage of the bibliometric data extracted from Harzing's Publish or Perish. The scope of this study did not include the monitoring and evaluation phase of the PD approach; however, it is envisaged that the longitudinal nature of the digital datasets would provide an opportunity to monitor and evaluate the mid-to-long-term effects of PD scaling interventions, without the additional costs of field data collection.

- **Additional predictors of PD performance:** Big data sources enabled us to not only identify PDs, but also to understand PDs in new ways that would not have been possible using traditional data sources. It still holds true that the "Discovering predictors of PD performance" stage relies mainly on primary data collection; however, in the Egypt case and in the Mexico project, big and open data were used to identify some predictors of PD performance that complemented predictors identified from interviews and surveys. For instance, in the Mexico project we explored how nighttime light imagery can be used to understand if the outdoor lighting brightness of the PD public spaces is higher than that in non PD spaces. In the Egypt case study, we used the publications of the researchers and applied content analysis and machine learning techniques to identify paper-extrinsic (e.g. number of pages) and paper-intrinsic predictors (e.g. topics covered) along with predictors related to the publication outlets (e.g. where do they publish their research) to characterize the research corpus of PD researchers.
- **Expanding the scope of positive deviance:** There has been a domain and geographic concentration in the application of the PD approach which has to date been predominantly applied in public health and in rural environments (Albanna & Heeks 2019). DPPD enabled the introduction of PD in new domains where it has not been applied before e.g. deforestation, crime control and rangeland management. And thanks to its large geographical cover, DPPD enabled us to look into the collective performance of individuals (e.g. vegetation biomass of villages and communities) and into the outcomes of spaces and physical environments (e.g. gender-based violence in public spaces). DPPD made it possible to look at units of analysis that are beyond individuals (e.g. villages, groups and communities) and we were able to look into supra-individual factors that are driving outperformance and could be transferred and amplified e.g.

development interventions, policies, governance schemes, etc. This holistic understanding of factors driving performance would inform the design of nuanced interventions that take into account the individual's behaviour as part of a larger system, which increases the effectiveness and the contextual fit of such interventions.

- **Analysis by-products:** When applying the DPPD method, we have to ensure the 'sameness' of the observed units when it comes to their resources and limitations. This is why DPPD adds the 'homogenous grouping' step prior to the PD identification, where positive deviants are identified within homologues (groups of units that have a similar context). In most of the projects, big environmental data complemented socio-demographic and economic data in creating those homologues. For instance, in the Indonesia and Niger projects, we cannot assess the performance of agricultural systems in isolation of their biophysical and environmental conditions such as rainfall, temperature and soil. Big environmental data enabled us to include those factors along with other contextual factors to create agroecological homologues. Through discussions with government and development stakeholders, it turned out that those homologues are of value regardless of their specific use in DPPD as these groupings will enable these stakeholders to better capture and understand the complexity of the agricultural systems which could inform their sampling methods. Such analytical by-products only derive from DPPD and would not have been produced under a conventional PD approach. Additionally, because it can build performance measures of relevance at national and subnational scale, DPPD allows identification and investigation not merely of positive deviants but also of negative deviants and other "bad performers". This can provide further guidance for national and subnational priority-setting of development interventions and resource allocation.
- **Leveraging machine learning:** DPPD enabled the introduction of machine learning techniques in the PD approach, since such techniques work best with large datasets. We touched on its benefit in the homogenous grouping stage, where we used various

unsupervised clustering methods (e.g. K means, PAM[44], HCPC[45]) to divide our study population into homologues having similar structural characteristics. Clustering techniques were also used to identify different types of outperformance by clustering observations based on multiple performance indicators, similar to the profiling presented in the Egypt case, where PDs were further classified into three subgroups of positive deviance, which later on can be studied separately and predictors specific to each group can be identified. This clustering would benefit from the large sample sizes afforded by DPPD and would be less applicable using the relatively smaller samples of the conventional PD approach. For PD identification, machine learning techniques were used in both the univariate analysis and the multivariate analysis. In the former, we applied unsupervised dispersion-based and model-based outlier detection techniques on the performance measure of the Indonesia project to identify outliers. As for the latter, machine learning models that require vast amounts of data (e.g. XG Boost, Neural net) were used to predict performance based on a set of contextual variables as independent variables to identify context-aware positive deviants based on the residuals (observed - predicted). In most cases, such models achieved a higher predictive power than traditional parametric regression models (e.g. linear regression). And higher predictive power implies better explained variance, hence more accurate residuals and a higher likelihood that the identified positive deviants are true. The performance of such models would be compromised if applied on a small number of observations, as found in the conventional PD samples. Furthermore, machine learning techniques were used in PD understanding, as in the Egypt case, where an unsupervised topic modelling technique

[44] Partitioning around medoids (PAM) clustering was used to cluster urban blocks in the Mexico project using a mix of categorical and continuous variables.

[45] Hierarchal clustering using principal components (HCPC) was used to cluster villages in the Indonesia project using time series rainfall and temperature data.

(LDA[46]) was used to identify the topics on which researchers work by analysing the abstracts of their publications.

6.3 What are the challenges of using big data in the PD approach?

Having presented the perceived benefits of using big data in the PD approach, I list here some of the challenges we faced when applying DPPD across the different pilot projects. Challenges ranged from human resource scarcity, data accessibility and reliability to privacy issues and hidden additional costs. Those challenges emerged from the application of the DPPD method across the case studies and projects presented in Chapters Three, Four and Five.

- **Flexibility with project scope is needed:** DPPD works best in domains and countries where data landscapes are rich with accessible and relevant contextual data. And, most importantly, where a digitally recorded or observed measure can be used to capture the outcomes of the observed population. However, these requirements are not always met, especially in developing countries where data landscapes are incomplete and, in some cases, entirely non-inexistent for specific development issues. Hence, in DPPD, the project design is largely dependent on secondary data availability and data should not be viewed as a mere technical aspect that can be accounted for once the project implementation has started. This requires flexibility and creativity in adjusting project scope based on data accessibility, reliability and the available resources, while maintaining a focus on the problem i.e. the development issue to be addressed. This flexibility could mean adjusting the study population sample, limiting it to areas with sufficient data coverage, or where outcomes could be measured digitally with confidence. For instance, while there is a well-established relationship between grain yields and remotely sensed vegetation indices (Ferencz et al. 2004), this relationship is less established with other types of crops (e.g. potatoes). Hence, in the Niger and Indonesia

[46] Latent Dirichlet Allocation (LDA) was applied over the entire corpus to identify topics and calculate the probability distribution across topics for each document. The next step was to compare the distribution of topics in the corpus pf positive deviants versus the distribution of topics in the corpus of non-positive deviants.

projects, our study populations were limited to grain-growing regions and seasons. Creativity may be needed in the selection of proxy performance measures that could indirectly and reliably measure the outcomes of the study population. For example, in the Somalia project, rangeland health was selected as a proxy measure of pastoral livelihoods resilience as it was much easier to measure rangeland biomass health using remote sensing vegetation indices, than it was to enumerate livestock numbers using costly high resolution satellite imagery.

- **While DPPD reduces data collection costs, there are hidden additional costs:** Mapping out data landscapes early on is necessary to assess the feasibility of the DPPD method. However, there is a tension between the relative rapidity of the mapping exercise, and the lengthy process of securing access to this data, and ensuring its reliability. Put simply, the mapping exercise may indicate that data exists, but it does not guarantee it can be used. Additionally, very specific know-how is needed for this assessment, which may not be easily accessible and attainable in international development organizations while projects are still in the design phase. The challenge then is to maintain a focus on the problem, as time and resources are spent on efforts to secure access to and validate data, without the guarantee that these will succeed; while properly considering whether to reframe the problem according to whatever data is readily accessible. In the Somalia project, the initial idea was to enumerate livestock numbers using high resolution satellite imagery and control for transhumance or livestock movement using call details records (CDRs). It took us three months and very specific domain expertise to realize that the needed high resolution imagery would be very costly to obtain, with no guarantee that it will accurately capture the livestock numbers. And that accessing CDRs would require a very long data access approval process with the telco company, also without guarantee that we would finally get this data. Moreover, fieldwork is still needed in DPPD: for a smaller sample since DPPD narrows down the qualitative search space, but there are still primary data collection costs in stage three despite their significant reduction in stage two thanks to the reliance on already existing non-traditional data. There are also costs related to the analysis of non-traditional data. Here, the time and cost of applying DPPD will vary largely depending on the in-house data analysis capabilities. If they exist there will not be additional costs needed to hire

consultants or delays to get them on board, which would in turn increase the cost of the method.

- **Data accessibility:** Applying the DPPD method does not just require access to any data, but often needs access to very specific and (sometimes) sensitive data with high spatial and temporal resolution. While in the initial stages of the pilot projects data might seem to be readily-available and usable at first glance, this may turn out to be less suitable for the analysis due to issues with granularity and/or coverage. Highly granular data on the other hand might not be publicly available, and might be costly to obtain (e.g. very high resolution satellite imagery) and often only accessible through partnerships with data holders. Most of the DPPD projects piloted in this study experienced difficulties in accessing usable data to varying degrees. There were difficulties related to delays in getting access, permissions to get access and money needed to purchase the data. These challenges obliged adjustment in project designs, in some cases applying a data-first approach using the data that was available, accessible, and reliable instead of the data that had been envisaged as an ideal fit for the project. In the Mexico project there are two sources that can be used to enumerate crimes in a certain area: the crime investigation reports which are found on the Open Data Portal of Mexico City, and the 911 calls that require access from the Control, Command, Communication, Computer and Quality Centre in the State of Mexico. It took us almost six months to get access to the 911 calls and, to avoid delaying the project until access was granted, we decided to move forward with the readily available crime reports data and build and refine the model using this data until it was possible to augment our crime report data with the 911 data.
- **Big data is de-contextualized data:** While celebrated for its potentials, big data is usually a by-product of already existing processes and it is very likely that it was collected for a purpose other than the purpose of the PD analysis we are aiming to conduct. Hence, big data suffers from a potential 'loss of context' which decreases the meaning and value that can be extracted from it (Bornakke & Due 2018). Additionally, it is often thin as its huge volume extends along only one or few dimensions of the observed performance. When applying DPPD we realized the importance of complementing big data sources with traditional secondary data sources to cover the contextual gaps that big data is not able to capture. This way we can form a fuller picture of the contextual realities of the observed population and, hence, are able to identify true PDs i.e. outperformers relative

to their respective context and not just in an absolute sense. Such contextual data is the basis of our performance estimates that are compared with the observed performance and the difference (i.e. residuals) is used to identify positive deviants. Hence, contextual data is of outmost importance because if it is outdated or inaccurate, it can lead to wrong predicted estimates, and thus to false outliers or positive deviants. In the Mexico project, we used the marginalization index, the daily incoming trips and the population density to undertake the homogenous grouping. At the beginning of the PD analysis, we were able to extract those variables from a 2020 census survey, except for the marginalization index which was only available for the year 2010. Halfway through the analysis, we received feedback from a domain expert that there is a risk in using an outdated marginalization index, which led us to re-engineer creation of the marginalization index using variables[47] from the 2020 census. We then redid the PD analysis on this basis. The more recent marginalization index changed the results of our initial grouping by almost 30% and led to the discovery of new PDs and the exclusion of PDs identified earlier. This is just an example of how PD identification is sensitive to changes in contextual data. An additional challenge of using this complementary data is that it should overlap both spatially and temporally with the big data-derived performance measure while being granular enough to be linked to individual units of observation. The geographical coverage and spatial resolution requirements are usually met in census data, but not fully met in the temporal coverage, as censuses take place every 5 or 10 years depending on the type of census (e.g. agricultural, population and housing, etc.) and the country where it is collected (less frequent in developing countries). Sample surveys have a quicker turnaround time, hence can provide more recent data. However, they usually cover only part of the population and are not as granular as the census surveys; hence the observation unit (e.g. village) might be linked to contextual data at a higher aggregation level (e.g. district).

- **Niche know-how:** Identifying a digital performance measure for DPPD that can be used as a reliable replacement for field measurement is a challenging task. It usually involves

[47] The variables used to create the marginalization index (e.g. illiterate population, employed population, housing conditions) for 2010 were documented. The same variables were extracted from the 2020 census data to create the 2020 marginalization index.

the use of proxy measures from big data that can indirectly measure the outcomes of the study population while making sure that potential contextual confounds are identified and controlled. While it is generally accepted that this type of approach holds important value for sustainable development, it is not yet mainstreamed in operational efforts (Njuguna & McSharry 2017). Deriving proxies from big data requires not only knowledge about the domain and the data sources that can be used, but also a fair understanding of the local context and conditions. Additionally, big data analytics is different from traditional statistical analysis, as it requires specific techniques and analytical methods to transform data into value in a time- and cost-effective manner. One of the biggest obstacles of this study in the projects reported here was the scarcity of local advanced analytics experts that have experience in the domain data. For instance, in the Niger project we needed to process huge volumes of high resolution Sentinel data to extract the SAVI vegetation index. The local expert we hired had sufficient domain know-how but her remote sensing data expertise was limited to offline GIS software (e.g. QGIS). The limited internet connectivity in Niger made it hard to download Sentinel data in a timely manner: it would have taken us months to download the imagery covering the study population due to its high resolution (10m). So, we were left with two options, either to use lower resolution satellite imagery such as MODIS (250m) that would take a few days to download but would compromise the granularity of the performance measure, or to hire another consultant who could process the Sentinel data on the cloud using the Google Earth engine without having to download it. We ended up choosing the second alternative in order not to risk identifying positive deviants at such coarse resolution. Finding experts that have this very specific combination of local, domain and data know-how at the project design phase was a very challenging yet crucial task. Especially so given that very few international agencies have it in-house, and that hiring external consultants for this kind of expertise, when it is not guaranteed that there will be a viable digital measure to apply DPPD, could be risky and costly.

- **Privacy concerns:** The small sample size and grounded nature of the conventional PD method, makes it possible to get the consent of the data subjects before measuring their outcomes for PD identification. Hence, it allows for having individuals as the units of analysis. This is not the case in DPPD, which starts with digital data covering large populations. Yet, for responsible big data research, necessary policies and regulations on

the protection of privacy must be respected (Pulse 2012; Zook et al. 2017). Hence, in the application of DPPD, the focus on individuals must be temporarily set aside, with the focus for PD identification being turned to higher-level units of analysis such as communities or geographic units which would not compromise the privacy of individuals. With the PDs among these higher-level units identified, they can then be the focus for fieldwork, looking for individual PDs within the higher-level unit and directly obtaining their consent for data gathering. Taking this route and aggregating performance to a higher level is based on the assumption that means are sensitive to outliers and that their values will be skewed towards the outlier value signalling the existence of PDs within them. For example, in the Niger project, what we measured was the agricultural outcomes of cereal growing villages across the Sahelian region in searching for outperforming villages. The following step was to visit those villages in order to identify farmer plots that are driving this outperformance and to obtain the consent of the farmers to collect data about their uncommon practices and strategies. However, there are exceptions where DPPD can have individuals as both the unit of observation (i.e. the level where outcomes are measured) and the unit of analysis (i.e. the level where we find solutions and pitch findings), when an implicit consent can be assumed. This was seen in the research performance case where researchers allowed for publicly available citation indices (e.g. Google Scholar) that are tracking their research outputs. An alternative would be when a third party has already collected the consent of the data subjects and is authorised to indicate their consent, making it possible to reuse this data.

- **Spurious correlation**: To be able to derive recommendations from the PD determination stage, it is important to identify causal mechanisms (i.e. processes or pathways through which an outcome is brought to being). Yet it has long been recognised that such derivation on the basis quantitative analysis may be fraught with problems, particularly the danger of mistaking correlation for causation (Lucas 1976). The DPPD approach makes major use of quantitative approaches to identify statistically-verified determinants of positive deviance and there is therefore some danger of falling into this trap, notwithstanding the statistical significance of the identified factors, In the best case, this could lead to ineffective or wasteful recommendations. In the worst case, it could do harm through the unintended negative consequences of inappropriate “solutions”. As discussed above, DPPD seeks to correct for this through qualitative, field-based validation

of causal mechanisms. Nonetheless, and particularly in its early stages before fieldwork, it may be appropriate to make use of quantitative techniques that can seek to cut through the correlation—causation conunudrum. Pearl (2003) for example, outlined methods of more reliable statistical causal inference using a series of methods that could be incorporated into DPPD, including non-parametric structural equations, graphical models, and counterfactual or "potential outcome" analysis (Pearl 2003). Using such methods for causal inferences might mitigate dangers of confusing correlation with causation, and thus lead to a more robust understanding of the casual mechanisms that underpin the complex systems within which DPPD-relevant problems often sit.

6.4 Contribution

This thesis makes substantial and original methodological contributions to knowledge. The principal methodological contribution is the iterative development of the data-powered positive deviance (DPPD) method through its application in the action research projects presented in Chapter Five; Chapter Five; projects which were part of a programme of GIZ-/UNDP-funded action research arising solely from the publication of the Chapter Two paper. The method combines traditional and non-traditional data to identify and understand deviance at geographical and temporal scales that were not possible using the conventional PD method. The systematic literature review presented in Chapter Two outlined challenges facing the application of the positive deviance approach in development, such as the time and cost of data collection (Lapping et al. 2002; Marsh et al. 2004; Felt 2011), cost efficacy and the relative rarity of positive deviants (Marsh et al. 2004), in addition to the narrow domain and geographic scope of PD applications in developing countries (Albanna & Heeks 2018). The novel DPPD method developed as part of this thesis has demonstrated its ability to circumvent some of those challenges while also contributing to better ways of identifying positive deviance such as the ability to observe performance over time and investigate units of analysis at aggregation levels that are higher than those normally used in the conventional method.

Previous literature has also documented the need to develop more context- and domain-specific frameworks to operationalize the adoption of PD (Singhal et al., 2010; Herington and van de Fliert 2017). This thesis contributes to that through the development of a DPPD

analytical framework that is specific to the scientific research domain as presented in Chapter Three, and a second framework that is specific to the agricultural domain as shown in Chapter Four. Those frameworks provide guidance on how to apply and replicate the DPPD method in those domains specifically, but also offer a more general approach that can be replicated in other domains. This thesis also contributes to the field of big data for development. Previous literature in this field has noted that one of the main challenges of the "big data paradigm", is not the large amount of data as much as the ability to analyse this data for intelligent decision making (Hilbert 2013). DPPD provides an analytical framework that could enable the transformation of digital information into knowledge that informs intelligent decisions. Previous literature also suggests the need to develop methodologies to characterize and detect socio-economic anomalies in context (Pulse 2012). DPPD serves us one such method, since the positive deviants identified in the method are in statistical terms context-aware outliers or anomalies, and they can be identified at both ends i.e. positive and negative deviants. Additionally, it has been well documented in a number of studies that there is a need to integrate thick data with big data to extract meaning and value from big data and to rescue it from potential context loss (Smets and Lievens 2018; Ang 2019). The DPPD method does this by providing a new way in which to systematically integrate insights from both types of data, where big data is used to identify who is doing well, while thick data is used to explain why they are doing well.

Alongside the methodological contributions, the application of the DPPD method contributed new analytical findings in each of the sectors in which it was applied. For example, in the Egypt research publication case study presented in Chapter Three, new hypotheses were generated and then tested. In this case study, publication-level and individual-level predictors of research performance were investigated to explain factors contributing to the outperformance of a small group of global South researchers. A mixed methods approach was employed which started with an inductive inquiry (interviews and topic modelling) to generate hypotheses from a small sample of positive deviants followed by a deductive inquiry (surveys and publication analysis) to identify significant differences between PDs and non-PDs. Through this study, and this approach, it was possible to identify predictors of PDs in a global South context that had not been identified in previous studies, and to provide pointers to ways of overcoming challenges specific to Southern researchers.

These predictors related to publication and co-authorship strategies, international collaboration and capacity building. Similar new analytical insights were generated into rice-growing in Indonesia, as summarised in Chapter Four.

6.5 Recommendations for Policy, Practice and Future Research Direction

6.5.1 Development Policy and Practice

The PD approach can have significant implications for development policy and practice. PD implies that the solutions needed for development are already existing within communities and organizations. As these solutions are locally sourced, they are less vulnerable to social rejection and they are more sustainable since they reduce the reliance on external aid and expertise**.** PD follows an approach of appreciative inquiry which starts by asking the question: what is working well and how can it be amplified for a greater effect? Instead of asking the question: what is wrong and how it can be fixed? (Ochieng 2007).

In the PD approach, what usually happens is that solutions adopted by positive deviants are used to inform interventions that rely on community-based participatory approaches to amplify and mobilize those solutions. The solutions are usually practices, strategies and attitudes at the level of individuals. The initial findings across the projects implemented in this study suggest that the DPPD method can identify and amplify not just solutions at the level of individuals but, more often, would identify both community-level and policy-level interventions. This is because the DPPD method - due to its large spatial coverage - can employ a more holistic lens that takes into account both individual and supra-individual factors to understand the complex forces at play behind a 'solution'. This is also an artefact of DPPD's aggregation level which can go beyond individuals to observe - say - entire cities or regions and see how for example certain policies, strategies and systemic changes affect the performance of those units differently. This is different from the conventional PD approach which is placed in a more 'controlled' environment and using a much smaller sample where variation in performance might indeed be attributed only to individual-level factors.

From this perspective, then, DPPD has two particular policy/practice implications that relate to scope and scale of application of positive deviance. First, DPPD should enable development policy-makers and practitioners to more-readily make use of PD methods in their work; having demonstrated application of positive deviance in a wider variety of sectors and to a wider variety of development challenges than was previously the case. As availability of big and other digital data sources grows, so too will this scope. Second, DPPD expands the levels at which positive deviance is applicable. While many development organizations have interests in individual behaviour and outcomes, many others are focused at a higher level. DPPD opens up for them at least the possibility that the PD approach will now be more-readily applicable to their interests.

There is also a growing interest by policy makers in evidence of what interventions result in successful outcomes and why (Sutherland et al. 2004; Kapos et al. 2008; Post & Geldmann 2018). DPPD can be utilised as a tool for both impact assessment and evaluation in programmatic positive deviance, which is concerned with understanding why a few positive deviants respond to a development intervention programme better than their peers who are targeted by the same intervention (Albanna & Heeks, 2019). Although not directly tested in the rounds of projects reported here, the same big data-derived performance measure that was used to identify positive deviants can be used to identify those exceptional responders[48] within a known intervention population. Interrogating the contextual factors that contribute to an exceptional outcome can make those exceptional responders valuable sources of information that can inform the design of more effective interventions.

In sum, then, DPPD provides a new methodology and new opportunities within positive deviance which itself – as noted in Chapter 1 – can be seen as one component within the overall set of best practices approaches to international development. It is hoped that greater recognition of the relation between DPPD and best practices could lead to further uptake, or at least further exploration, of DPPD among those who see themselves as part of the best

[48] To avoid reaching false conclusions due to an erroneous result or a chance outlier, quasi-experimental or experimental study designs should be adopted before information obtained through this approach is used to inform policy (Post and Geldmann, 2018).

practices approach. This might particularly be relevant to those programmes and organisations which have been attracted to quantitative approaches to development such as the Bill and Melinda Gates Foundation (Fejerskov 2015). Recognition of this connection does, though, also identify a future challenge for DPPD – the wider the scope of its application (i.e. with higher-level interventions), the greater the dangers of mismatch between context and solutions. As international development more broadly has done, moving from the terminology and concept of "best practice" to that of "best fit" may therefore be helpful (Ramalingam et al. 2014).

6.5.2 Data Policy and Practice

DPPD presents a new method for the extraction of value from big data, and can join the repertoire of analytical approaches used by data scientists, especially in the field of data for development. The method is still in its relative infancy and its further adoption and application by practitioners in the field would contribute to its development and the reliability of its findings. After its application across the different domains and projects outlined in this study, earth observation (EO) data proved to be one of the most viable big data sources that can be used in the DPPD method. Thanks to the recent advances in its temporal, spatial and spectral resolution, EO data is able to measure and observe the outcomes of natural and built environments at a global scale with longitudinal coverage. Additionally, such data can often be acquired without any administrative restrictions, making EO-derived performance measures ideal for applying DPPD and reducing the data collection time and cost. Four out of the six projects reported in this study used EO data: three to measure vegetation biomass and one to observe forest change. It can therefore form a particular focus for further adoption of DPPD.

As noted earlier, there are data-specific challenges related to the use of the DPPD method such as data access, lack of the right know-how and infrastructural limitations. Those challenges manifest particularly in Southern countries, which triggers the need for data policies that are not specific to the DPPD method alone, but which are necessary for the viability of data-driven innovation in the global South more generally. First would be open data policies that could make development-related data more accessible and partnerships more attainable. So far, the attention of the open government and open data movements in

relation to such policies has been largely on high-income and middle-income countries with a skew towards Europe and the Americas, leaving out most low-income countries of the global South (Open Data Watch 2014),. This must change with resources allocated to promote the benefits of open data in developing countries which are now lagging in open data terms. Such resources could be invested, for example, in programmes targeting government and other public sector organizations that could demonstrate the value of opening their data to the public and/or which would mandate or in other ways encourage such opening. Such policies would need to go beyond supporting one-time release of data using in-house formats but support ongoing release of 'cleaned' data in widely-usable formats. There are plenty of exhortations for governments to open their data on the basis of claimed benefits such as greater transparency and accountability of government, greater equality in society, or economic value (Verhulst & Young 2017). However, in practice, these factors tend to be of limited importance to governments and may even, in the case of greater transparency of government actions, be a disincentive (Gonzalez-Zapata & Heeks 2015). What activists need to engage more with is the actual drivers (Huijboom & Van den Broek 2011; Toots et al. 2017). These include citizen demand, which can be stimulated and articulated by working with citizen groups and mass media; international reputation and national/international legislation which, for example, the Open Government Partnership and its membership programme has sought to promote; and the personal promotional incentives of appearing to be an innovative thought leader and practitioner. Alongside these drivers, there must also be a presence of key enablers to open data including digital and human infrastructure (Jetzek 2013). These issues apply more broadly to all forms of data that could be the basis for DPPD in developing countries, with a particular issue around data science capabilities (Joubert et al. 2021). In-house data science capabilities in the public sector are usually lacking and opening up this data creates an opportunity to make better use of it by data science capabilities available outside of government departments. Programmes can also target the private sector (e.g. telcos) to demonstrate the value of non-traditional data sources for development, and to incentivize making their internal data available without compromising the privacy of their users. General building of data science capabilities in the global South – addressing all stages of the data value chain from collecting and cleaning to processing and visualising data – will be required if DPPD is to be locally appropriated as a method for informing development policy and practice. That in turn will require the building of data

science education infrastructure: both human and technological. More general data-related infrastructural challenges, such as internet connectivity and storage capacities, might partly be addressed using infrastructure as a service (IaaS), cloud services such as Amazon web services, and open source software such as Google Earth engine (Luna et al. 2014). Use of these platforms can, in turn, be built into data science education and training.

As noted above, one of the key limitations of using big data is the potential loss of context, and context is specifically important in the PD approach because positive deviants should be identified relative to their context and not in an absolute sense. To mitigate the risks of big data-associated decontextualization, secondary traditional data sources were used extensively to integrate context into the data analysis. Hence, it is advised that practitioners would employ DPPD in data landscapes and domains that are relatively rich with readily available contextual data that can complement big data. This data should be recent data, that covers the study population and with a spatial resolution that can be linked to the project's unit of observation. Examples of such development-related, data-intensive sectors include but are not limited to: health, education and other public services. However, how granular, frequent and easily integrable and accessible this data is, would differ from one country to another, and would be particularly challenging in Southern countries. One way to overcome some of those issues is to present compelling, practical cases of how value can be extracted from such data by international actors (Taylor et al. 2014). Such actors with superior access to data and analytical capacity can demonstrate the means of and benefits achieved from analysing this data in their national contexts.

Also emerging from this study was the recommendation of involving the different stakeholders (public, private, non-profit and study subjects) and domain-specific data experts very early on when applying DPPD projects. Their involvement ensures: 1) the correct identification of the problem; 2) what is considered outperformance and if it exists; 3) what outcome measures can be used and how they can be extracted from non-traditional data sources; 4) what usable data exists and how it can be accessed; and 5) what are potential confounds and contextual factors that have an influence on the performance measure. Furthermore, having the buy-in of the different stakeholders at the very beginning guarantees, to some extent, the adoption and amplification of findings from the PD inquiry later on; be it through a community intervention or a policy intervention.

6.5.3 Next Steps for DPPD Projects

This study focused on the first three stages of the DPPD method, without presenting means on how to go about the last two stages of scaling practices and evaluating impact which were outside the scope of the current research. However, there is a clear and well-defined framework and process for the scaling out of PD-identified new practices, which would not differ much from DPPD-identified practices (Pascale et al. 2010). This process of diffusion is inside-out, with an emphasis on "doing" rather than "seeing" or "hearing". It is community-driven in the sense that the role of the expert lies in facilitating the mobilization of the solutions, by creating spaces where non-positive deviants can practice those solutions and learn about them from positive deviants. This PD process has been explicitly contrasted with and differentiated from approaches such as those based on "Diffusion of Innovations" theory (Rogers 2010) which are seen as more top-down and expert-driven (Singhal & Svenkerud 2009, Singhal 2011, Singha & Svenkerud 2018, Bhattacharya & Singh 2019, Dearing & Singhal 2020). Notwithstanding this, there is still potential for DPPD to incorporate ideas and practices from other approaches to scaling of new practices in development, especially considered the potential of DPPD for large-scale which might favour less dependence on community mobilization. For instance, elements derived from DOI theory, such as features of the innovation, the dissemination strategy, the alignment of the external environment with adoption of the innovation, and features of the adopting organizations, or users have been applied to help guide scaling in a few large-scale PD studies, such as the study by Bradley et al. (2009). However, the DOI ideas hold quite a specific place within the scale-out: "In such a scheme, hierarchical position of DOI will be subservient to PD, a change agent, a catalyst, and a leader. DOI chronicles the timeline, course, and profile of "followers" whether innovators or laggards." (Bhattacharya & Singh 2019). Similarly, ideas from models such as social network thresholds (Valente 1996) could find some place in informing the design of PD intervention scale-out by helping determine the proportion of adopters in the social system needed for an individual to adopt an innovation, and hence working out the "tipping point" of adoption level to which resources must particularly be addressed. Additionally, new digital technologies and tools can open up new opportunities for scaling DPPD-identified solutions. An example would be the use of targeted social media ads for PD-informed behavioural

change interventions in high-penetration contexts such as urban areas of developing countries with high levels of social media usage.

6.5.4 Future Research

In this thesis, the DPPD method has been developed iteratively through its application in a set of action research projects, and presented in a portable form that can be applied on other projects. However, there are multiple dimensions along which the scope of applying DPPD can be expanded. Firstly, domain expansion, with potential to apply DPPD in other domains and development challenges that were not explored in this study such as health, education, good governance, etc. Secondly, expanding the type of data used: other sources of big data that could capture the behaviours and outcomes of individuals and groups, such as mobile, IoT and social media data, have not been examined, and that merits further investigation. Thirdly, other units of analysis still need to be tested, at both higher and lower levels of aggregation. In particular, in would be valuable to explore examples of individual-level analysis where DPPD is feasible. Similarly, higher levels of aggregation such as regions and countries were not examined in this study, which prompts further research to test its applicability.

Additionally, in this study I only covered the first three stages of the DPPD method: problem-method fit, determining positive deviants, and discovering PD practices and strategies. More work is needed to assess the value and limitations of using big data in the last two stages of the DPPD method concerned with designing and implementing interventions, and monitoring and evaluating their effects on the intervention population. In both stages, the value from big data is not yet validated and further experimentation is needed. Finally, another methodological application that can still be explored is the use of non-traditional data in identifying exceptional responders i.e. those who respond significantly better to interventions in comparison to others in the intervention group.

In sum, DPPD is a novel method that combines traditional and non-traditional data sources to identify what works and why. It enables development practitioners to make evidence-informed decisions from the increasingly available non-traditional datasets, which is the natural step in the ongoing evolution from the "information age" to the "knowledge age"

(Hilbert 2013). The method I have developed is still at its operational infancy, holding both promises and challenges in its application in development. But I hope that the evidence provided in the case studies and action projects presented in this study will help it become a standard part of the data-for-development repertoire in the future.

References

Albanna, B. & Heeks, R. (2019) Positive deviance, big data, and development: A systematic literature review, *The Electronic Journal of Information Systems in Developing Countries*, 85(1), e12063.

Ang, Y. Y. (2019) *Integrating Big Data and Thick Data to Transform Public Services Delivery*. Washington DC: IBM Center for The Business of Government.

Bhattacharya, S., & Singh, A. (2019). Using the concepts of positive deviance, diffusion of innovation and normal curve for planning family and community level health interventions, *Journal of Family Medicine and Primary Care*, 8(2), 336

Bolton, D. K. & Friedl, M. A. (2013) Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics, *Agricultural and Forest Meteorology*, 173, 74–84.

Bornakke, T. & Due, B. L. (2018) Big–thick blending: A method for mixing analytical insights from big and thick data sources, *Big Data & Society*, 5(1), 205395171876502.

Dearing, J. W., & Singhal, A. (2020). New directions for diffusion of innovations research: Dissemination, implementation, and positive deviance, *Human Behavior and Emerging Technologies*, 2(4), 307-313.

Fejerskov, A. M. (2015). From unconventional to ordinary? The Bill and Melinda Gates Foundation and the homogenizing effects of international development cooperation. *Journal of International Development*, *27*(7), 1098-1112.

Felt, L. J. & Cody, M. (2011) *Present Promise, Future Potential: Positive Deviance and Complementary Theory,* Unpublished manuscript. Available at: http://www.laurelfelt.org/wp-content/uploads/2011/06/PositiveDeviance-CodyMayer.LaurelFelt.Quals_.May2011.pdf (Accessed: 21 September 2021).

Ferencz, C., Bognar, P., Lichtenberger, J., Hamar, D., Tarcsai, G., Timár, G., Molnár, G., Pásztor, S.Z., Steinbach, P., Székely, B. & Ferencz, O.E. (2004) Crop yield estimation by satellite remote sensing, *International Journal of Remote Sensing*, 25(20), 4113–4149.

Gonzalez-Zapata, F., & Heeks, R. (2015). The multiple meanings of open government data: Understanding different stakeholders and their perspectives. *Government Information Quarterly*, *32*(4), 441-452.

Herington, M. J. & van de Fliert, E. (2018) Positive deviance in theory and practice: A conceptual review, *Deviant Behavior*, 39(5), 664–678.

Hilbert, M. (2013) Big data for development: From information-to knowledge societies, *SSRN Electronic Journal*, 1–39.

Huijboom, N., & Van den Broek, T. (2011). Open data: an international comparison of strategies. *European Journal of ePractice*, *12*(1), 4-16.

Jetzek, T., Avital, M., & Bjørn-Andersen, N. (2013). Generating value from open government data. Paper presented at *Thirty Fourth International Conference on Information Systems*, Milan, 15-18 Dec.

Johnson, D. M. (2014) An assessment of pre-and within-season remotely sensed variables for forecasting corn and soybean yields in the United States, *Remote Sensing of Environment*, 141, 116-128.

Joubert, A., Murawski, M., & Bick, M. (2021). Measuring the big data readiness of developing countries–index development and its application to Africa. *Information Systems Frontiers*, advance online publication.

Kapos, V., Balmford, A., Aveling, R., Bubb, P., Carey, P., Entwistle, A., Hopkins, J., Mulliken, T., Safford, R., Stattersfield, A. & Walpole, M. (2008) Calibrating conservation: New tools for measuring success, *Conservation Letters*, 1(4), 155–164.

Lapping, K., Marsh, D.R., Rosenbaum, J., Swedberg, E., Sternin, J., Sternin, M. & Schroeder, D.G. (2002) The positive deviance approach: Challenges and opportunities for the future, *Food and Nutrition Bulletin*, 23(4_suppl_1), 128–135.

Lucas, R. (1976). Econometric policy evaluation: A critique. In Brunner, K. & Meltzer, A. (eds.), *The Phillips Curve and Labor Markets*, New York, NY: Elsevier, 19-46.

Luna, D. R., Mayan, J. C., García, M. J., Almerares, A. A. & Househ, M. (2014) Challenges and potential solutions for big data implementations in developing countries, *Yearbook of Medical Informatics*, 23(01), 36–41.

Marsh, D. R., Schroeder, D. G., Dearden, K. A., Sternin, J. & Sternin, M. (2004) The power of positive deviance, *British Medical Journal*, 329(7475), 1177–1179.

Mkhabela, M. S., Bullock, P., Raj, S., Wang, S. & Yang, Y. (2011) Crop yield forecasting on the Canadian prairies using MODIS NDVI data, *Agricultural and Forest Meteorology*, 151(3), 385–393.

Njuguna, C. & McSharry, P. (2017) Constructing spatiotemporal poverty indices from big data, *Journal of Business Research*, 70, 318–327.

Ochieng, C. M. O. (2007) Development through positive deviance and its implications for economic policy making and public administration in Africa: The case of Kenyan agricultural development, 1930-2005, *World Development*, 35(3), 454–479.

Open Data Watch (2014) *Overcoming Open Data Worries*. Available at: https://opendatawatch.com/blog/overcoming-open-data-worries/ (Accessed: 18 January 2022).

Pascale, R., Sternin, M. & Sternin, J. (2010). *The Power of Positive Deviance*. Boston, MA: Harvard Business Press.

Pearl, J. (2003). Statistics and causal inference: A review. *Test*, *12*(2), 281-345

Post, G. & Geldmann, J. (2018) Exceptional responders in conservation, *Conservation Biology*, 32(3), 576–583.

Pulse, U.N.G (2012) *Big Data for Development: Challenges & Opportunities*. New York: UN Global Pulse.

Ramalingam, B., Laric, M., & Primrose, J. (2014). *From Best Practice to Best Fit: Understanding and Navigating Wicked Problems in International Development.* London, UK: Overseas Development Institute.

Rogers, E. M. (2010). *Diffusion of innovations*. New York, United States: Simon and Schuster.

Singhal, A. (2011) Turning diffusion of innovation paradigm on its head: The positive deviance approach to social change, in *The Diffusion of Innovations*, A. Vishwanath & G. A. Barnett (eds). New York, NY: Peter Lang, 193–205.

Singhal, A., & Svenkerud, P. J. (2018). Diffusion of evidence-based interventions or practice-based positive deviations, *Journal of Development Communication*, 29(2).

Singhal, A., & Svenkerud, P. J. (2019). Flipping the diffusion of innovations paradigm: Embracing the positive deviance approach to social change, *Asia Pacific Media Educator*, 29(2), 151-163.

Smets, A. & Lievens, B. (2018) Human sensemaking in the smart city: A research approach merging big and thick data, *Ethnographic Praxis in Industry Conference Proceedings*, 2018(1), 179–194.

Sutherland, W. J., Pullin, A. S., Dolman, P. M. & Knight, T. M. (2004) The need for evidence-based conservation, *Trends in Ecology & Evolution*, 19(6), 305–308.

Taylor, L., Cowls, J., Schroeder, R. & Meyer, E. T. (2014) Big data and positive change in the developing world, *Policy & Internet*, 6(4), 418–444.

Toots, M., Mcbride, K., Kalvet, T., & Krimmer, R. (2017). Open data as enabler of public service co-creation: Exploring the drivers and barriers. In *2017 Conference for E-Democracy and Open Government (CeDEM)* (pp. 102-112). New York, NY: IEEE.

Tucker, C. J. & Sellers, P. J. (1986) Satellite remote sensing of primary production, *International Journal of Remote Sensing*, 7(11), 1395–1416.

Valente, T. W. (1996). Social network thresholds in the diffusion of innovations, *Social networks*, *18*(1), 69-89.

Verhulst, S. G., & Young, A. (2017). *Open Data in Developing Economies: Toward Building an Evidence Base on What Works and How*. Cape Town, South Africa: African Minds.

Zook, M., Barocas, S., Boyd, D., Crawford, K., Keller, E., Gangadharan, S.P., Goodman, A., Hollander, R., Koenig, B.A., Metcalf, J. & Narayanan, A. (2017) Ten simple rules for responsible big data research, *PLOS Computational Biology*, 13(3), e1005399.