学校代码：　10246
学号：15210690837

復旦大學

硕士学位论文
（专业学位）

公共部门效率和数据整合的非参数化方法。

Non-parametric approach to public sector efficiency and data integration.

院系：　管理学院

专业学位类别（领域）：　国际商务硕士

姓名：　José Juan Martín Quesada

指导教师：　张诚

完成日期：　2017 年 3 月 30 日

Table of Contents

1. ABSTRACT

世界各国政府都在紧张地加强预算，同时又刺激经济的增长。但由于可自由支配的支出和结构性支出的削减幅度受到经济和社会的限制，加上财政状况不断恶化，必须找到创造性工具来提高公共部门货物和服务的交付效率。

这些新型工具的关键在于大数据、机器学习和其他高级分析工具，这些工具依赖于健全的数据采集政策以及公共管理部门对政策的实施情况。在这种效率竞争中，政府以自身方式取得的边际改善可能意味着能节省数百万美元，创造成千上万个新的就业机会，甚至拯救生命。由于国家储蓄已经耗尽，宏观经济杠杆透支，根据其潜力，确立哪些工具有最大影响力并确立他们愿意对其进行多少投入将有利于政府的利益。因此，本研究的目的旨在探讨公共部门的数据整合是否确实能影响政府效率；如果能，那么具体是数据周期的哪一部分最重要，以至于吸引了政府的注意力，并使政府纷纷解囊。我们采纳了万维网基金会的开放数据晴雨表计划（Caulier-grice 等，2015），将“数据整合”从单个预测因素分成“准备”，“实施”和“影响”三类。

为了确定数据整合周期的哪部分对提高公共效率至关重要，我们首先分析后者，重点放在投入上。 这是为了区分有效的行政管理与高效的行政管理，这种区分建立在两个方面: 一方面是公民对他们从政府那得到的产出的评估，另一方面是从国家总支出来讲，提供所述货物的成本。随后，我们通过将数据整合分解为组件分数，并使用几种方法来探索这些分数与公共部门效率之间的关系，从而定义我们的预测因素。

我们假设，公共部门效率受数据整合的影响，这点在抽样的 77 个国家中得到验证以及三年份的数据。事实上，即使控制了像腐败水平或公民财富这样的地方特点，数据整合也被看作是效率的重要预测因素。多年来证明相关的变数为市民的电子线上参与工具(影响力)、强而有力的数据保护监管框架的存在(就绪度)、城市或地

方政府的开放数据举措的存在(就绪度)、企业和个人数据培训的存在(就绪度)、以及有关犯罪统计数据的大量数据（实施）。在精细度较小的层面，由多个数据类型的存在、批量可用性等定义的*实施*在大多数情况下证明是相关的，而在对国家所在地区进行会计时，准备度最为相关。我们运行层次线性模型，显示了当我们按区域进行分类，并且引入腐败作为控制变量时，*实现*和*影响*的重要性。我们使用*随机森林*和*有限混合模型*来确定自变量及其潜在类的相对重要性，并将它们逐一在表中加以定义。最后，主要部分分析将影响生产力的最重要因素简化为两部分: 1. 各国政府亲自积极参与数据社区（相对于仅基于数据发布的被动方式）以及 2. 政府的数据政策用于提高公共效率，而非帮助民间社会实现该效率。

Keywords:

公共部门效率, 数据, 政府, 效率, 公共政策, 公共行政, 管理, 大数据, 高级分析, 政府效率

Governments across the world are being pressed to consolidate budgets while stoking growth, but macroeconomic levers and social cohesion both risk exhaustion. In this context, a key tool for improving the efficiency of delivery of public sector goods and services is the integration of Big Data, machine learning and other advanced analytics tools into the operation of government, as well as the promoting of those technologies within the wider civil society. The purpose of this study is to explore whether the integration of data in the public sector does indeed have an impact on government efficiency; and if so, which parts exactly in the data cycle are the most important so that governments can direct their attention and resources to developing their expertise there.

We create an index to gauge public sector efficiency across a global sample of countries and confirm the effect of data integration in improving efficiency even after controlling for corruption, wealth and other variables through a logical succession Ordinary Least Square regressions, Hierarchical Linear Models, Random Forests, automatic and Gaussian Finite Mixture Models, as well as Principal Component Analysis. Additionally, we identify which concrete policies and government actions are most useful across countries and years in the pursuit of government efficiency.

Keywords:

Public sector efficiency, data, government performance, public policy, public administration, management, Big data, advanced analytics, government efficiency

2. INTRODUCTION

The end of 2016 and all of 2017 mark a time when a quadruple set of circumstances have the potential to greatly affect the public finances of governments worldwide, and those in Europe in particular: firstly, the Federal Reserve of the United States has ended its ultra-low interest rate policy in the sight of rising employment levels and higher levels of inflation closer to the long-term goal of yearly increases of 2% ("Decisions Regarding Monetary Policy Implementation - December 14, 2016," n.d.)

Secondly, the European Central Bank is only months away from commencing the tapering of its purchase of all types of assets, currently worth €80 billion a month, but set to be reduced to €60 billion monthly ("Monetary Policy Decisions - January 19, 2017," n.d.), with sovereign bonds having the potential of being disproportionately affected (Hale, 2017). Thirdly, the Western world, Japan and other relevant markets, are seeing a medium-cycle change in the trend of inflation due mainly to rising energy and commodity prices (Antonio M. Conti, Stefano Neri, 2017). Fourthly, the recently inaugurated administration in Washington has signaled its intention to increase spending in infrastructure and the military, as well as its intention to lower taxes. If these expectations remain large and consistent, they raise the prospect of billions in capital flowing into the United States (Giavazzi & Pagano, 1995). Additionally, the uncertainty created by the prospect of the United Kingdom leaving the European Union, and the terms of said departure, have depressed the value of the pound to the point where high-quality assets may seem more attractive to outside investors than those from their European counterparts.

The combined effect of these circumstances is twofold: for one, higher inflation and interest rates elsewhere risk transferring yield-maximizing investors' interest away from artificially-depressed returns in the Euro area and into other markets. A second effect is the shock in sovereign bonds interest rate caused by an outright retreat of one of the main investors in European bonds, the European Central Bank (ECB). While these figures are not published, Dutch bank Rabobank estimated the size of the European sovereign debt pool to stand at €7.5 trillion by mid-2016, while the full ECB asset-buying program comprises €1.6 trillion. A study from the ECB itself estimated the effect of the asset-purchase program to have resulted in a GDP-weighted reduction of sovereign yields in the Euro area of 77 basis points between September 2014 (with the implicit announcement of the program being priced-in by investors) and February 2015 (Santis, 2016). This impact is likely to have been larger when accounting for more recent dates, and the same study indicates that this impact was magnified in the case of the most highly indebted countries.

Therefore, the position of the most indebted countries relative to their GDP looms particularly delicate in a context where favorable financing conditions are about to wane. With this urgency in mind, governments have looked into reducing spending, increasing revenue and fostering growth. The macroeconomic implications of aligning these seemingly conflicting interests fall outside of the scope of this study. However, some consensus has arisen in the past few years highlighting the perils of largely indiscriminate budget cuts ad consolidation hurting aggregate demand proportionally more than they constrain public spending, and thus being futile (Figari et al., 2015). Since fiscal consolidation comes at the expense of social cohesion and, potentially, future growth, it is in governments' best interest to utilize "more disaggregated evidence to reach robust policy conclusions" (Figari et al., 2015).

It is in this context of prolonged fiscal consolidation, when most obvious fiscal levers have been exhausted and monetary policy is all but fully stretched to its potential, that governments can benefit most from streamlining expenses, reducing error and refining policy decisions. Key amongst the tools to help achieve this are Big Data and Advanced Analytics techniques. These were once novel in the private sector, but as their implementation in enterprises has increased, their potential in the public sector remains largely underutilized.

A seminal study conducted by McKinsey & Company (Manyika et al., 2011) identified €150 to €300 billion in value which Europe's 23 largest governments could create over 10 years in 3 main areas: Operational Efficiency Savings (value: €120 to €200 billion), Reduction in Fraud and Error (value: €7 to €30 billion) and Increases in Tax Collection (value: €25 to €100 billion). A revisitation of the issue in 2016 by the same institution found that only 10% to 20% of this total potential value had been captured in the previous 5 years (Henke et al., 2016). If realized, the value captured could amount to productivity gains of 0.5% per year across Europe.

The purpose of this text is to prove or disprove that government productivity is indeed affected positively by the degree of availability and integration of data (both defined below) in their functioning, once singularities such as their income level, HDI scores, government corruption levels or region are controlled for. The secondary purpose of this text is to identify which parts of data integration are most valuable to government productivity, such as "Data Readiness", "Data Implementation" and "Data impact".

Before continuing, some concepts ought to be clarified:

1. Public sector productivity: generally defined as unit of output per unit of input.

$$P_t = \sum_{i=1}^{n} \frac{O_{it}}{I_{it}} \quad (1)$$

Where productivity *(P)* at any given point in time *t* is the result of dividing the output *(O)* of every *i* area measured by the inputs *(I)* in that area.

The very notion of public sector productivity, and the idea that it could and should be measured, is a fairly novel concept. For years, it was common practice to assume that public sector productivity was zero, and so the output created by government was calculated by simply equating it to the inputs used. This practice has now become outdated, and "outcome" has been accepted to comprise administrative, educational and health results, as well as "the quality of public infrastructure, the support of the rule of law" and a the creation of opportunities which create "a level playing-field in a market economy" (Head, 1970). Most analyses will also feature some approximation to the measurement of the Musgravian tasks for government (allocation, redistribution and stabilization) as well as the measurement of input and output efficiency (Afonso, Schuknecht, & Tanzi, 2005).

Therefore, regardless of what i areas are included in the measurement of productivity and the metrics *(M)* that make up said area, if $P_t = f(M_k)$, then any changes in government productivity must come from a change in the underlying metrics such that:

$$\Delta P_t = \sum_{i=1}^{n} \frac{\partial f}{\partial M_k} \Delta M_k \quad (2)$$

2. Big data: definitions vary, but most revolve around the concept of large volumes of data that are produced routinely by organizations and are too complex for standard software packages to process (Mayer-Schönberger, 2013).

3. Availability and integration of data: for data to be available it has to *exist*, be accessible *in bulk* and *in machine readable formats*. Additionally, governments must have policies in place to make use of such data (*Readiness),* to put those policies in practice (*Implementation)* and to do so in effective initiatives (*Impact)* (Caulier-grice et al., 2015).

3. LITERATURE REVIEW

We start with a review of the previous work on measuring government efficiency, and specifically, on the convenience of utilizing different metrics to gauge input and output levels. We then assess prior studies on the impact of Big Data on the public sector at different administrative levels, the areas of most potential, as well as some case studies.

3.1 GOVERNMENT EFFICIENCY

3.1.1 MACROECONOMIC APPROACH TO GOVERNMENT EFFICIENCY MEASUREMENT

A number of attempts have been carried out to try to measure public sector productivity. The difficulty in most of the cases is that in order to measure public sector effectiveness and public sector performance, the outcomes of government activity need to be quantified, even when there are no market prices or information on the volume of "transactions" to put a value on said outcomes (for example, in the value of the continued provision of public security). Past attempts have ranged from "whole of government" productivity approaches (Schnabel, de Kam, Kuhry, & Pommer, 2004) to individual country studies across time. The most comprehensive early instance may have been the one executed by the European Central Bank, which undertook an international comparison seeking to separate the concepts of inputs and outputs in the public sector based on a number of performance indicators (Afonso et al., 2005). Prior to this study, most attempts equated the value of public sector output to the inputs received in terms of budget utilized and sometimes also labor, ignoring changes in productivity (Atkinson, 2005). Instead, the aforementioned study selected 7 indicators divided into 2 groups: "Musgravian" (as per Richard Musgrave's contribution to classifying government activity into resource allocation, income redistribution and macroeconomic stabilization (Musgrave, 1939)) and "Opportunity" indicators, which englobe metrics indicative of how well governments promote societal well-being and prosperity. The latter measured in terms of educational attainment, state of public infrastructure, volume and quality of healthcare and effectiveness of administrative tasks. This approach allowed for comparable inter-country results, as the metrics were not affected by different measurement techniques used by each country assessed.

Finally, we focus on the attempt by the World Bank to appraise the quality of governance in the *Worldwide Governance Indicators*. A yearly study across a panel of countries and throughout time, starting in 1990. This account does not measure public sector productivity strictly speaking because the scores they provide can only be used for the measurement of government output, not inputs. We will examine this study in more detail further below with the rest of the data, as the information provided by their survey is an integral part of the analysis in this thesis.

3.1.2 DIRECT ESTIMATION APPROACH TO GOVERNMENT EFFICIENCY MEASUREMENT

The ECB study covered only OECD countries, and it did so on the basis of pondering the importance of macroeconomic figures which might not reflect structural circumstances (i.e. does not take into account the initial stock a country possesses of any one particular metric, such as pre-existing wealth or infrastructure). The publication of the "Atkinson Review" (Atkinson, 2005) provided the guidelines that the UK Centre for the Measurement of Government Activity has been using ever since to evaluate total public service output, input and productivity through direct measurements. These reports itemize output and input for the activities which consume most government resources, such as healthcare and education, and uses the older input = output convention for the rest.

Many of the Atkinson review guidelines have become the de facto standard in the granular measurement of public sector efficiency. It introduced concepts such as healthcare outputs being assessed by specialized metrics like the promotion of healthier living, concretized on primary data like “decreased of deaths related to smoking”. These quantity metrics are then adjusted for quality of delivery, to arrive at final outcome figures. This approach, while being exhaustive and allowing for comparisons within same-sector departments and over time, does not allow for cross-country evaluations. The scope of this type of analysis is also outside of the means employed for this thesis.

3.2 INTEGRATION OF DATA IN THE PUBLIC SECTOR

As mentioned above, one of the most important initial studies on the topic was the one conducted by the McKinsey Global Institute in 2011, where they outlined the biggest profit pools for the public sector to benefit from by exploiting data in several novel ways or by doing so more extensively. While this report focused on high-level macroeconomic estimations of potential impact, further studies have given us a glimpse into concrete applications of big data at the regional and even local level.

At the regional level, we describe here a case example of potential applications that have been proved to be perfectly actionable with current technology. Analysts at bank BBVA together with researchers from the United Nations Global Pulse conducted an experiment where they used credit card purchase and ATM transaction data before and after a natural disaster to glean insights about levels of preparedness and patterns in the recovery of different regions in the aftermath of said event (Martínez et al., 2016). Certain communities where shown to stock up on essentials more extensively than others, and some communities were seen returning to “normal” transaction levels sooner than others. These data had the immediate effect of helping direct state relief efforts to areas where it was most needed.

At the local level, we find the example of the application of clustering techniques like those employed in seismology in order to discern whether crime statistics could reveal “contagion” patterns whereby local burglary events prompted subsequent events around the same area (Mohler, Short, Brantingham, Schoenberg, & Tita, 2011). These insights could hardly have been obtained without the large-scale modelling of previously inaccessible Los Angeles Police Department data.

4. SOURCES OF DATA FOR THE ANALYSIS

We first list the main sources of data for the measurement of government productivity, divided into public sector output and public sector input. Later we describe the sources employed to measure the degree of integration of data in government operations globally.

4.1 PRODUCTIVITY DATA

4.1.1 PUBLIC SECTOR OUTPUT DATA

There are very few recent studies which cover global public sector productivity. The choice confronted was between using the first, mostly outdated attempts such as the one undertaken by the ECB; opting for in-depth country analyses which are limited to a few individual regions such as the UK, and which do not follow comparable standards; or building a database from scratch. Given the level of detail and granularity needed for these productivity metrics to be actually meaningful, the last option was deemed outside of the scope of this study. Instead, the choice was made to construct a new index that would allow for inter-country comparison using separately available sources for output and input. This is because there are datasets with enough level of detail and granularity to allow for the separate assessment of both public sector output and input. Specifically, we choose the *Worldwide Governance Indicators* (WGI) published yearly by the World Bank to act as proxies for the evaluation of output in equation (1). The downside is that because these metrics were not built to be used jointly like we intend to, combining them will yield an index that is only useful to compare between the countries included, but which hold no significance as a standalone figure.

The WGI are an appropriate choice because they aggregate many individual output metrics from disparate data sources and they combine them into a comparable indicator evaluating the quality of governance expressed in terms of quality perceptions in healthcare, education and infrastructure. This is done by using an unobserved components model (UCM), in which the premise is that “each of the individual data sources provides an imperfect signal of some deeper underlying notion of governance that is difficult to observe directly”. This means that, as users of the individual sources, we face a “signal extraction problem”. This extraction problem is addressed by the UCM model (Kaufmann, Kraay, & Mastruzzi, 2011). For each component of governance, we assume that “the observed score of country j on indicator k, y_{jk}, can be written as a linear function of unobserved governance in country j, g_j, and a disturbance term, ε_{jk}” such that:

$$y_{jk} = \alpha_k + \beta_k(g_j + \varepsilon_{jk}) \quad (3)$$

Where α_k and β_k are parameters which “map unobserved governance in country j, g_j

into the observed data from source k, y_{jk}". These observed scores y_{jk} are then rescaled to have common units, and the final individual aggregated score per country is simply a weighted average of all the rescaled observed scores. Larger weights are given to more informative data sources (the ones which have the smallest variance of error ε_{jk}over the total variance).

The WGI are divided into six sections, including topics such as rule of law and control of corruption. While these concerns are also very much influenced by government behavior, we have restricted our interpretation of output to the *Government Effectiveness* section. This section delves mainly on the quality of public services (education and healthcare chief amongst them), as well as the quality of infrastructure and the civil service. The reasoning behind employing this section only as a proxy for government output is twofold. Firstly, areas where government involvement can be important such as civic involvement and accountability, are not necessarily relevant to the measurement of government efficiency. Political stability and security are also not the exclusive dominion of governments, but arguably they are the result of a series of political choices in the political constitution of states which escape the government of the moment. Lastly, we have decided to treat corruption levels as a control variable rather than as an outcome of government action that would not be very indicative of productivity in and of itself. The second reason why we use only *Government Effectiveness*, even if we disregard all concerns cited, is that there had to be a direct link between outputs and inputs for the relationship to be meaningful. Given that there is limited quantifiable data on inputs for areas such as political stability or rule of law, we decided to follow the Atkinson's Review recommendation to restrict the *output* areas considered to those that make up the majority of government expenditure (*inputs* in equation 1) and are easiest to measure; education, healthcare and civil infrastructure.

4.1.2 PUBLIC SECTOR INPUT DATA

Once the choice was made to use the WGI data for outputs, corresponding metrics to the areas measured by WGI had to be used for inputs. The World Development Indicators from the World Bank database were deemed the most reliable and comprehensive dataset. The process to gather a balanced dataset occurred following the following steps. Later stages were not reached when a previous stage had already yielded the necessary data:

1. Retrieval of healthcare and education general government expenditures as a percentage of total government expenditure, 2004 to 2015: only government spending is considered, which leaves out institutions run entirely on private capital. Since these private institutions can constitute a significant percentage of total healthcare and education spending, we opted to express government spending in these areas as a percentage of total government spending instead of the more common "as a percentage of GDP". This has the additional advantage of eliminating distortions due to extreme outliers in the income distribution. That is, since total government spending also includes money originated from international transfers, very poor countries would have exceedingly high expenditures as a percentage of GDP, because their spending is also partly funded by international transfers (mostly aid). Conversely, nominally rich countries do not spend a constant percentage of their GDP in public services, so their spending might have appeared exceedingly small as a percentage of their GDP, but not as a percentage of their total spending. Government expenditures includes all levels of government; national, regional and local. The years before and after 2004 and 2015 were not considered due to incompleteness. Thus, healthcare spending would be expressed as:

$$Healthcare\ spending_t = \sum_{i=1}^{n} \frac{H_{it}}{S_{kt}} \quad (4)$$

Where Healthcare Spending *(H)* in period t for all health-related activities i, divided over Total Government Spending *(S)* for all government activities k equals the percentage sought.

2. For the considerable number of countries and years for which the abovementioned data were not available, we instead obtained the figures as percentages of GDP and then proceeded to divide them by general government spending as a percentage of GDP figures, so that all data were expressed in terms of percentages over total government spending. For healthcare:

$$Healthcare\ spending_t = \left[\left(\sum_{i=1}^{n} \frac{H_{it}}{GDP}\right) / \frac{S_{kt}}{GDP}\right] \times 100 \quad (5)$$

3. For the cases where the previous two stages had left single years devoid of datum, an average of the previous three years' figures was employed.

4. When despite all the previous stages a number of years still lacked data, these data were searched for individually. Sources included OECD country statistics, IMF financial statistics, UNESCO and World Health Organization data, Eurostat, national statistical offices data and, ultimately, press releases or news articles.

5. Finally, for the very few remaining cases where some consecutive years were still missing information, a three-year average was calculated for the latest year missing, and that same figure was used as a proxy for the missing adjacent data. The reasoning being that this way at least the denominator of equation (1) would remain constant in the face of incertitude over the real figures, but the numerator would still change and provide a part of the real variance.

Additionally, the overall percentages of government expenditure over GDP were also conserved to examine correlation with WGI scores, although these figures included items not measured by the Government Effectiveness section of WGI, such as military expenditure, foreign representation and many others.

Finally, figures for Gross Formation of Fixed Capital (GFFC) were retained. This metric can be equated to the net investment in physical assets (disregarding financial assets) during a year in any given country, which is the same thing as the gross investment in physical assets minus gross disposal, disregarding depreciation and operational expenses. This figure is a part of the national accounts of most countries, and comprises the net investment carried out by the private sector (business and households), and government. Therefore, an attempt was made to obtain the figures for *private* Gross Formation of Fixed Capital per country so that, after subtracting from the total GFFC we would arrive at a figure for net *public* investment in physical assets which could roughly be equated to public investment in civil infrastructure. Unfortunately, the disaggregation of GFFC data per private/public sector was not available for many countries. As a result, the unbalanced dataset was used only subsidiarily for additional insights and in the end the calculated net public FFC figures varied so widely that they could not be used to establish correlations and were discarded.

4.2 DATA INTEGRATION IN THE PUBLIC SECTOR

Comprehensive, globally available information on the integration of data in the public sector is harder to obtain than public productivity metrics. That is because the technologies on which Big Data and Advanced Analytics techniques in general are based are recent. There are many possible proxies, such as the level of development of IT infrastructure, the level of IT education of the population in general, the expenditure on R&D, etc. All of these are too generic, so a more granular source had to be found.

Our final analysis has made use of many elements from the Open Data Barometer published by the World Wide Web Foundation (Caulier-grice et al., 2015), as well as the United Nations E-Government Survey (United Nations Department of Economic and Social Affairs, 2016). The latter ranks countries in terms of their telecommunications infrastructure, human capital, online services available to citizenry, as well as the use by civil society of online government services ("*e-participation*").

As mentioned, the Open Data Barometer (ODB) has been our principal source of information. It is a worldwide assessment of Open Data policies, their enactment and results in governments across the world. Wherever the information could not be obtained from government databases, the World Wide Web Foundation conducted expert surveys which underwent significant back and forth to ensure that results were as reflective of reality as possible and that, whenever personal opinion was involved (e.g. in "To what extent" types of questions) a similar gauge was employed everywhere. We have used many of the sub-indicators in the original barometer, then decomposed the sub-indicators into each of their components, added some additional metrics (in the table below called "*Code"*) from outside sources and eliminated others which might add collinearity with our control variables or are simply not relevant (such as "*Political Freedoms and Civil Liberties*" for example). Then we standardized all of those figures versus their country peers in the metric group and recalculated the sub-indexes for a new comprehensive final score. The resulting variables, and their description are the ones listed below:

Sub-index	**Component**	**Code**	**Description**
readiness	government_policy	RE1	To what extent is there a well-defined open data policy and/or strategy in the country?
readiness	government_policy	RE2	To what extent is there a consistent (open) data management and publication approach?
readiness	government_policy	RE3	To what extent is there a robust legal or regulatory framework for protection of personal data in this country?

readiness	government_action	RE4	To what extent does the country have a functioning right-to-information law?
readiness	government_action	RE5	To what extent is there a well-resourced open government data initiative in this country?
readiness	government_action	RE6	To what extent are civil society and information technology professionals engaging with the government regarding open data?
readiness	regulatory_+_civil	RE7	To what extent is government directly supporting a culture of innovation with open data through competitions, grants or other support?
readiness	regulatory_+_civil	RE8	To what extent are city or regional governments running their own open data initiatives?
readiness	regulatory_+_civil	RE9	To what extent is training available for individuals or businesses wishing to increase their skills or build businesses to use open data?
readiness	Busin. & entrepren.	RE11	Government online services index
implementation	innovation	IM1	Map Data
implementation	innovation	IM2	Land ownership data
implementation	innovation	IM3	Detailed census data
implementation	innovation	IM4	Detailed government budget
implementation	innovation	IM5	Detailed data on government spend
implementation	social_policy	IM6	Company register
implementation	social_policy	IM7	Legislation
implementation	social_policy	IM8	Public transport timetables
implementation	social_policy	IM9	International trade data
implementation	social_policy	IM10	Health sector performance
implementation	accountability	IM11	Primary or secondary education performance data
implementation	accountability	IM12	Crime statistics
implementation	accountability	IM13	National environment statistics
implementation	accountability	IM14	National election results
implementation	accountability	IM15	Public contracts
implementation	dataset_assessment	Each implementation score is the sum of all	Is the dataset open?
implementation	dataset_assessment		Does the data exist?
implementation	dataset_assessment		Is it available online from government in any form?
implementation	dataset_assessment		Is the dataset provided in machine-readable formats?

implementation	dataset_assessment	these elemnts per score	Is the machine-readable data available in bulk?
implementation	dataset_assessment		Is the dataset available free of charge?
implementation	dataset_assessment		Is the data openly licensed?
implementation	dataset_assessment		Is the dataset up to date?
implementation	dataset_assessment		Is the publication of the dataset sustainable?
implementation	dataset_assessment		Was it easy to find information about this dataset?
implementation	dataset_assessment		Are (linked) data URIs provided for key elements of the data?
impact	political	PA1	Government E-participation index
impact	political	PA2	To what extent has open data had a noticeable impact on increasing government efficiency and effectiveness?
impact	political	PA3	To what extent has open data had a noticeable impact on increasing transparency and accountability in the country?
impact	social	PA4	To what extent has open data had a noticeable impact on environmental sustainability in the country?
impact	social	PA5	To what extent has open data had a noticeable impact on increasing the inclusion of marginalised groups in policy making and accessing government services?
impact	economic	PA6	To what extent has open data had a noticeable positive impact on the economy?
impact	economic	PA7	To what extent are entrepreneurs successfully using open data to build new businesses in the country?

Figure 1 Scores for data integration in the public sector. Sources: Web Foundation, World Economic Forum, Freedom House, United Nations, World Bank

The scores for Implementation have two components: firstly, there are fifteen categories of data generated by, or utilized by governments. These categories cover everything from land registry data to crime statistics. Secondly, there are degrees of implementation, ranging from the very existence of the data to how up to date it is. Each of the fifteen categories of government data is measured on the eleven implementation degrees for a final category score. These category scores will subsequently be added back together for the formation of the *Implementation* sub-index score.

Metrics were standardized in the following manner. For example, for metric *RE1* and *i* country:

$$re1 = \left\{ \left[{\left(RE1_i - MIN(RE1)\right)} \middle/ {\left(MAX(RE1) - MIN(RE1)\right)} \right] \right\} \times 100 \quad (6)$$

To obtain a final score per sub-index we initially followed the same procedure than in the ODB. This method is based on a series of weighted averages which we will omit here for brevity, but can be found in the ODB methodology section.

The result is set of 3 sub-indexes which we can use either to construct a final index or to explore which of the aspects of government adoption of data carries the most weight in their productivity results.

The *Readiness* scores reflect government attitude and action towards enabling the integration of data in their work, and also the measures being undertaken to extend the usage of data by businesses and other segments of society. *Implementation* reflects progress in the generation of open data sources for the most important areas where government-collected information can be employed; this progress being measured on the basis of the data existing, being accessible, provided in machine-readable form, for free, in bulk, etc.

Impact is a special case in that it partly measures what we aim to quantify with our productivity ratios. That is, assessing whether government policies and actions have had a noticeable effect on efficiency. We will take special care in checking for correlation and assigning a smaller weight if we include *Impact* together with *Readiness* and *Implementation* in a comprehensive index, but for now we leave this information in the database, as it may serve as a complement to *Implementation* to discern how far have governments gone from formulating policy to actually implementing it.

4.3 CONTROL VARIABLES

We have considered several control variables to account for eventualities like the fact that more developed countries may generally feature more efficient public sectors. Consequently, we introduce variables such as *Region* (East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, Sub-Saharan Africa), *HDI Rank* (Very High, High, Medium, Low) and *Income* (High income, Upper-middle-income, Lower-middle-income, Low income) following World Bank standard classifications.

HDI is a composite measure featuring education, health and income scores. Given the fact that we will measure the first two as part of our government *Output*, we mostly omit HDI as a control variable and instead focus on *Income* and *Region* amongst others.
Besides their wealth and location, the other significant country characteristic that could significantly affect productivity is broadly defined by "*rule of law*" (Sala-I-Martin, 1997). Once decomposed into sub-factors, we discard those whose effect will be accounted for in government *Output* and are left with a measure of corruption. In this respect, we opted to choose the most widely reputed and comprehensive index, the Corruption Perceptions Index published by Transparency International (*Corruption Perceptions Index*, 2016), a study which merges the analyses of other reputed institutions in gauging the perception of corruption levels in countries as measured by local experts and analysts.

5. HYPOTHESES AND METHODOLOGY

5.1 HYPOTHESES

We will attempt to prove the following complementary hypotheses:

> HYPOTHESIS 1A (H1A): *Data integration is positively correlated with public sector efficiency.*
>
> HYPOTHESIS 1B (H1B): *Readiness is the most important element of data integration for public sector efficiency.*

We follow a non-parametric approach to our model, where the relationship between the different independent variables and the dependent variables will be explored with several exploratory models, increasing the level of variable complexity by incorporating more and more of the countries' attributes and changing the way they are grouped.

We start, however, simply by checking normality and the possible correlation between our variables. These variables include the ones mentioned thus far, but also a simple ratio for *Productivity* which we will use as our dependent variable henceforth, called sP15 (for year 2015, sP14 for 2014 and so on). sP15 is simply a standardized ratio of government output, as defined by Government Effectiveness in the WGI dataset, divided by the combined healthcare and education expenditure over total government expenditure (this variable is called ED.HE.2015 for 2015 and so on).

$$P_t = \frac{WGI_t}{ED.HE_t} \equiv \sum_{i=1}^{n} \frac{O_{it}}{I_{it}} \quad (7)$$

For normality we use the Shapiro-Wilk test through the *nortest* package and *shapiro.test* command in R. Our data shows sP15, sP14 and sP13 (the initial years we are going to consider) to be normal for α = 0.05. This conclusion should, however, be taken with discretion, as the dataset is not large enough to draw definitive conclusions, for example, see the uneven histogram for productivity in 2013:

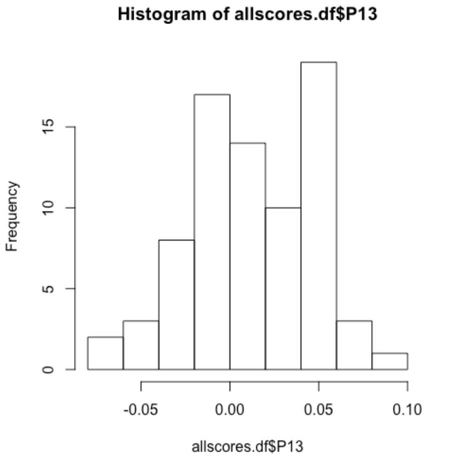


Figure 2 Distribution of productivity figures in 2013

The Anderson-Darling test, run through the *ad.test* command, assumes a certain distribution a priori unlike Saphiro-Wilk. This test seems to show more conclusively that certain years' productivity ratios are indeed not distributed normally. But the bell shape of the histogram and our dearth of balanced datasets for more countries (which in any case do not number exponentially more than the 77 countries studied) lead us to believe that should the sample be larger, the Central Limit Theorem would indeed apply. More importantly, we will check and correct for heteroscedasticity later on when our model is built.

We now establish what the correlations are between several of the independent variables:

Pair		**Description**	**R^2**	**p-value of correl.**
Income.2015	ED.HE.2015	Income - Expense	0.1375	0.01244
Income.2015	COR15	Income - Corruption	0.541	2.321e-12
ED.HE.2015	COR15	Expense - Corruption	0.04466	0.06505
Gov.Exp2015	COR15	Total exp - Corruption	0.1562	0.0003754
Gov.Exp2015	Income.2015	Total exp – Income	0.2495	0.000101
Region	ED.HE.2015	Region – Expenditure	0.196	0.01548
Region	COR15	Region – Corruption	0.4094	1.196e-06
Region	Gov.Exp2015	Region – Total exp	0.3572	1.799e-05

Figure 3 Correlations table

Most of the strongest correlations are expected. The higher the income level the higher the score in the absence of corruption index (where a 100 score means no corruption). If we treat income as a 4-tiered factor such as we described above, all levels would detract approximately 30 points from the corruption index score versus the default (High-income) which seems to suggest that only truly rich countries see a significantly lower level of corruption (and vice versa). This is consistent with previous research (Mauro, 1995) suggesting corruption slows down investment and growth and so it is likely to be correlated with lower relative levels of wealth.

There also seems to be a statistically significant, if small, correlation between government expenditure (both total and including only Education and Healthcare) and corruption. Again, this is supported by the findings of previous studies (Goel & Nelson, 1998) suggesting the size of government is indeed correlated with higher levels of corruption measured by convicted officials, at least at least at the state and local level. Corruption is also strongly inversely correlated with productivity, as can be seen in figure 5.

Geography is also very clearly correlated with expenditure. Unsurprisingly, European regions appear to dedicate on average 20 points of their GDP more in total expenditure than the smallest spenders, South Asian nations. Regarding corruption, North Americans score 40 to 45 points (over 100) better than South Asian and Sub-Saharan nations, respectively the best and two worst performing regions.

We will take special care in the treatment of the strongest of these correlations when incorporating them as control variables or factors in any form so as to avoid collinearity and preserve parsimony. In the following page, we can see box-and whisker plots representing the maximum and minimum values, first and third quartiles, and outliers per each type of classification.

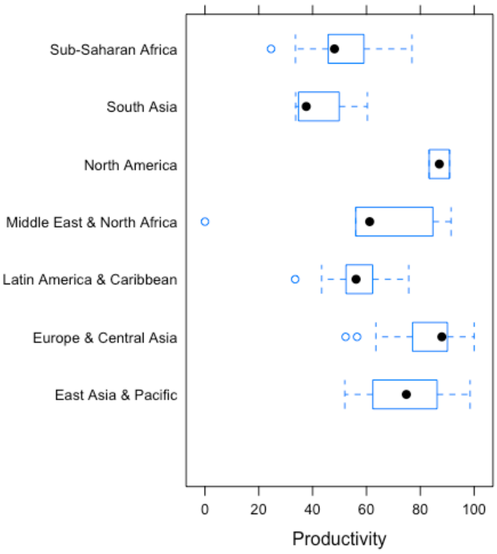


Figure 4 Box-and-whisker plot Productivity-Regions

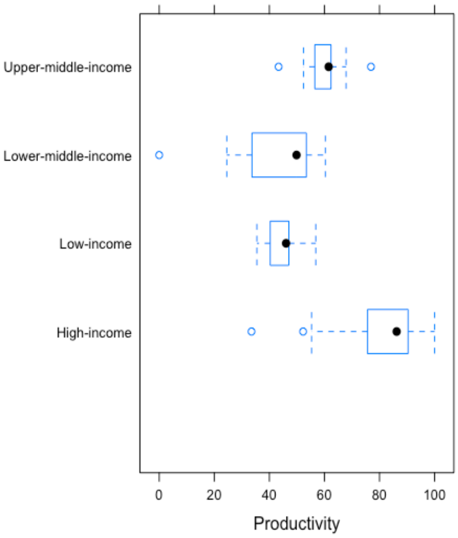


Figure 5 Box-and-whisker plot Productivity-Income

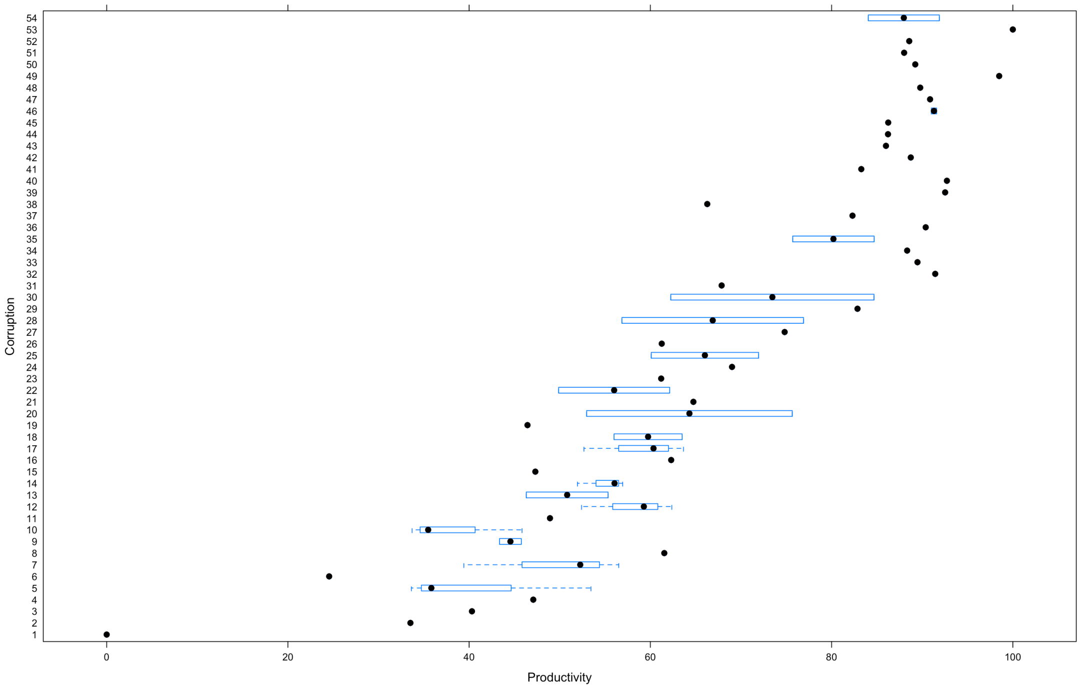


Figure 5 Box-and-whisker plot Productivity-Corruption

5.2 WORK PLAN

The logic we will follow as we explore different models hereon is based on the following steps: first, the most basic linear relationships suggest that indeed a composite score of data integration has a significant impact in productivity even when accounting for corruption levels or income. We then decompose the composite score into three sub-indexes to see which carry more weight in productivity changes, controlling both for region and corruption levels.

Because these linear relationships might behave differently for different income groups or regions, we regress the three sub-indexes in Hierarchical Linear Models that, after 200,000 iterations each, converge to account for countries belonging to different groups of wealth or geographic location.

The next step will be applying Finite Mixture Models to uncover patterns of latent factors combining from the observable elements to form "invisible" distributions. We let an automatic iterative process determine the number of necessary factors from a complete subset of all 32 basic metrics that make up the sub-indexes. We then force the algorithm to assume underlying clustering into 2 or 3 Gaussian models so as to explore which variables truly matter for a country to belong with its peers or not.

We also have decided to run 10,000 iterations of Random Forest instances that gives us an insight into the relative importance of each of the total 36 variables in explaining Productivity. This process is conducted with and without *Corruption* and *Income* so as to avoid the distortion of their great weights. The last steps include applying Principal Component Analysis to ascertain two large principal components, and applying all acquired knowledge into a final simple model incorporating interaction effects.

Ordinary Least Squares - Model 1

A simple ordinary least squares regression between productivity in 2015 and a composite score made up of all scores (*Readiness*, *Implementation* and *Impact*) finds a significant correlation with a strong R^2 of 0.556. Because our productivity index has no interpretable scale to economic data, both productivity and the composite scores have been standardized for comparability. An increase of 1 point in the composite scores results in a 0.68-point improvement in productivity. For this first model, the composite score has been calculated granting the same weights to each component has determined by the ODB.

$$Productivity = 39 + 0.68Composite + \varepsilon \quad \textbf{(Model 1.1)}$$

Model 1

	Dependent variable:
	Productivity 2015
Composite data scores	0.680^{***}
	(0.070)
Constant	39.376^{***}
	(3.124)
R^2	0.556
Adjusted R^2	0.551
Residual Std. Error	13.820 (df = 75)
F Statistic	94.093^{***} (df = 1; 75)
Note:	*p<0.1; **p<0.05; ***p<0.01

But evidently, we aspire to control for relevant categories, let us explore the relationship of some variables to *Productivity* so that we can then control for the most relevant ones.

Pair		Description	R^2	p-value of correlation
sP15	HDI.Rank.15	Productivity - HDI	0.7395	2.2e-16
sP15	Region	Productivity - Region	0.492	8.647e-09
sP15	Income.2015	Productivity - Region	0.638	4.317e-16
sP15	COR15	Productivity - Corrupt	0.8026	2.2e-16

Figure 6 Control variable correlations to productivity

Notwithstanding the strong correlation of Productivity to HDI, this relationship carries great overlap in the terms they measure, as explained above, so we discard it. We favor *Income* and *Corruption* as our control variables, bypassing the effect of their strong correlation by ignoring one when favoring other, applying hierarchical linear models that account for countries belonging to "*Rich*" or *"Corrupt"* groups, and exploring interaction effects between both.

$$Productivity = 21 + 0.16Composite + Corruption + \varepsilon \quad \textbf{(Model 1.2)}$$

The R^2 of this relationship is much larger at 0.8165, and both independent variables are still significant. However, their correlation is also high at 0.5548 so we will explore later better ways to incorporate these dimensions.

Ordinary Least Squares - Model 2

The effect of both the composite data scores and income levels are significant and behave as expected. Not only do progressively higher levels of income positively affect government productivity, but so does their integration of data. Even if the effect is not as pronounced as in model 1 it is still large and significant, as an improvement of 1 point in the composite scores impact productivity by one third of a point even after controlling for income. The exception of Lower-middle income levels subtracting more productivity than Low-income levels is more likely to signal to lesser availability of data for the least developed countries than to a non-linear relationship.

$$Productivity = 63 + 0.341Composite - 23Low\ Income - 27\ Lower\ middle\ income - 13Upper\ middle\ income + \ \varepsilon \quad \textbf{(Model 2)}$$

Model 2

	Dependent variable:
	Productivity 2015
Composite data scores	0.341^{***}
	(0.082)
Low-income	-23.005^{***}
	(5.166)
Lower-middle-income	-27.046^{***}
	(4.596)
Upper-middle-income	-13.639^{***}
	(3.830)
Constant	62.988^{***}
	(4.818)
R^2	0.708
Adjusted R^2	0.692
Residual Std. Error	11.439 (df = 72)
F Statistic	43.700^{***} (df = 4; 72)
Note:	*p<0.1; **p<0.05; ***p<0.01

The interaction effects between income levels and composite data scores are not significant in any instance.

Ordinary Least Squares - Model 3

When we disaggregate the composite data scores into their three components we see that all three of them lose their significance (for simplicity here Income has been made into a dummy variable that takes value 1 when the country belongs to the high-income group but the results hold for the 4-level factor too). This holds true even when we treat income as a first grouping factor in a hierarchical linear model. However, when we eliminate both *Implementation* and *Impact*, we see that *Readiness* has practically the same explanatory power than the composite index by itself.

$$Productivity = 41 + 0.191Readiness + 0.185Implementation + 0.027Impact + 18High\ income + \ \varepsilon \ \textbf{(Model 3)}$$

Model 3

	Dependent variable:
	Productivity 2015
Readiness	0.191
	(0.167)
Implementation	0.185
	(0.170)
Impact	0.027
	(0.150)
rich1	18.397^{***}
	(3.781)
Constant	40.984^{***}
	(3.135)
R^2	0.667
Adjusted R^2	0.649
Residual Std. Error	12.219 (df = 72)
F Statistic	36.072^{***} (df = 4; 72)
Note:	*p<0.1; **p<0.05; ***p<0.01

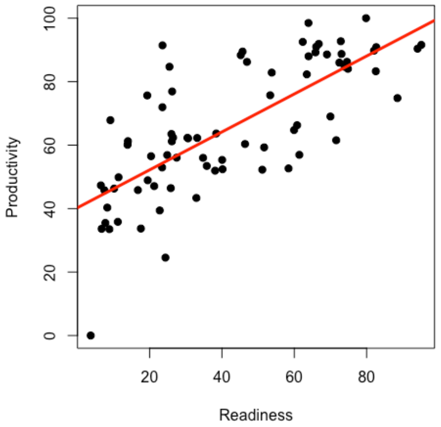


Figure 8 Readiness plotted against productivity

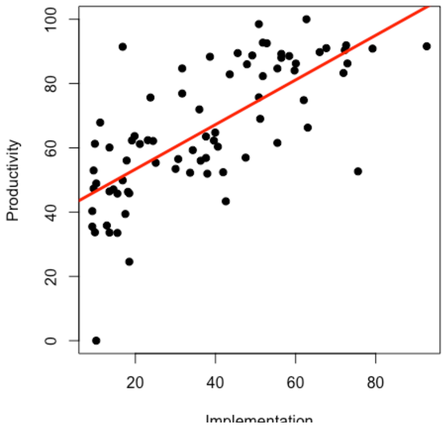


Figure 7 Implementation plotted against productivity

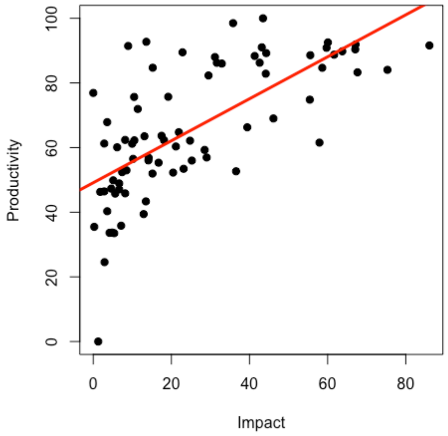


Figure 9 Impact plotted against productivity

This lack of relevancy for predictors other than *Readiness* is susceptible to respond to several explanations, some of which may have interesting implications for policy:

1. Most *Readiness* questions veer on the subject of governments having defined clear data-preparedness policies and procedures at different administration levels. Therefore, we must consider the possibility that this sub-index is the most relevant in explaining productivity because it is inherently indicative of how well organized and structured a government already is. That is, rather than help a government function more efficiently, data *Readiness* is indicative of how efficient that government already is.

2. *Implementation* is made up of 15 data categories with 11 implementation degrees each, only one of which is “The data for this category *exists*”. Therefore, once a government has collected the data, having standardized, machine-readable, easily-accessible bulk data might be helpful, but not crucial to overall efficiency. A further study of this topic could contrast the value for a country of increasing productivity by a certain amount versus the cost of proceeding with costly administration computer system upgrading and standardization programs that would go beyond having datasets that merely *exist*.

3. *Impact* explores the perceived success of implemented policies and procedures. This success relates to several categories which are also contained in our productivity index trough their inclusion in *government effectiveness* (the nominator in equation 7). However, not all *Impact* categories must necessarily carry the same weight in productivity, while currently our sub-index weighs impact on the economy or new businesses with the same weight as impact on inclusiveness, which while tremendously important for society as a whole, might explain government productivity changes less powerfully.

To address the issues of fixed weighs obscuring the real importance of scores in points 2 and 3 above, and to obtain insights beyond the constrains of the sub-indexes we have built, we will apply Principal Component Analysis (PCA) to the 32 individual scores which make up the sub-indexes, as well as Finite Mixture Models.

Before that, however, let us comment on the results of a Hierarchical Linear Model (HLM) approach.

Hierarchical Linear Models - Model 4

We examine the behavior of the sub-indexes under both *Income* and *Region* categorizations through HLM using the *lmer* tool in R. The lack of volume combined with the suspect irrelevancy of some of the sub-indexes means that even after converging at 200,000 iterations neither the random nor the fixed effect coefficients can be generalized due to lack of significance.

$$Productivity = Readiness + Implementation + Impact + (Readiness + Implementation + Impact \mid Income) + \varepsilon \quad \textbf{Model 4.1}$$

Random effects

Groups	*Name*	*Variance*	*Std.Dev*	*Corr*		
Income.2015	(Intercept)	2.850e+02	16.8834			
	Readiness	7.294e-03	0.0854	-0.81		
	Implementation	8.148e-02	0.2855	-0.03	-0.56	
	Impact	2.225e-01	0.4717	-1.00	0.86	-0.06
Residual	1.082e+02	10.4001				

Fixed effects

	Estimate	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(Intercept)	46.63842	8.97457	2.93600	5.197	0.0146
Readiness	0.17786	0.15002	14.42900	1.186	0.2549
Implementation	0.03019	0.21731	4.00400	0.139	0.8962
Impact	0.37633	0.28647	1.06800	1.314	0.4037

Correlation of Fixed Effects

	(Intr)	*Rednss*	*Implmn*
Readiness	-0.302		
Implementtn	-0.139	-0.500	
Impact	-0.719	-0.067	-0.107

The most visible coefficients are, of course, the income intercepts. All correlations are negative except *Impact* with *Implementation*.

Random effects coefficients	*(Intercept)*	*Readiness*	*Implementation*	*Impact*
Low-income	37.70279	0.21209333	0.04796553	0.62311759
Lower-middle-income	28.59070	0.26831032	-0.05486511	0.89242694
Upper-middle-income	63.12529	0.15215341	-0.21515562	-0.04799251
High-income	57.13487	0.07889384	0.34282310	0.03777290

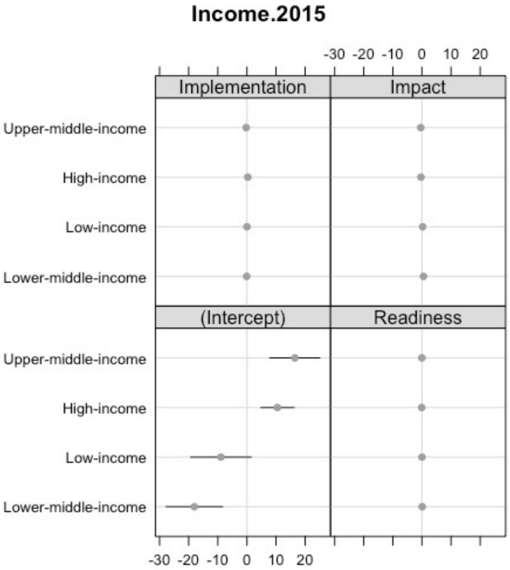


Fig. 10 Random effects plot for each income level

The random effect coefficients for *Readiness* are consistently positive, while the other two sub-indexes present mixed results depending on the income group. This might hint at a non-linear relationship or simply to a lack of data for poorer countries, as the coefficients for High-income countries are all consistently positive. Alternatively, another explanation might have to do with higher *Implementation* scores affecting productivity negatively because corrupt practices mean that a bigger availability of data results in more opportunities for graft. We will explore this with a model below.

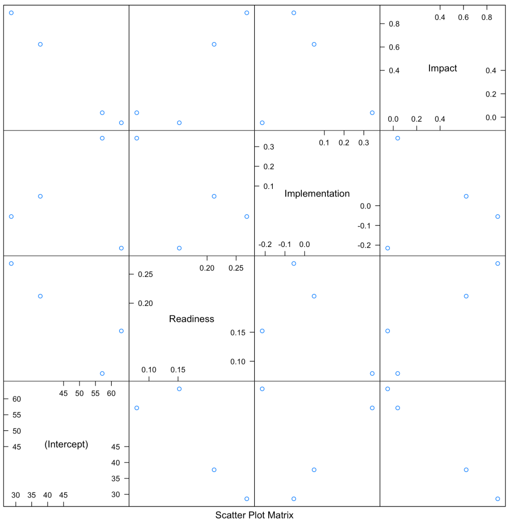


Figure 11 Coefficients plot

Let us try to re-categorize using regions instead of income levels.

$$Productivity = Readiness + Implementation + Impact + (Readiness + Implementation + Impact \mid Region) + \varepsilon \quad \textbf{Model 4.2}$$

Random effects

Groups	*Name*	*Variance*	*Std.Dev*	*Corr*		
Region	(Intercept)	7.134e+00	2.67098			
	Readiness	1.206e-02	0.10983	-1.00		
	Implementation	6.606e-02	0.25702	1.00	-1.00	
	Impact	3.138e-03	0.05601	-1.00	1.00	-1.0
Residual	1.625e+02	12.74820				

Fixed effects	*Estimate*	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(Intercept)	39.33366	4.09407	7.77700	9.607	1.39e-05
Readiness	0.35409	0.17907	18.9440	1.977	0.0627
Implementation	0.30289	0.22099	8.17900	1.371	0.2069
Impact	-0.04531	0.16674	22.7750	-0.272	0.7883

Correlation of Fixed Effects

	(Intr)	*Rednss*	*Implmn*
Readiness	-0.332		
Implementtn	-0.250	-0.600	
Impact	0.371	-0.436	-0.344

With all correlations being equal to |1|, this method of classification must be discarded.

Random effect coefficients	*(Intercept)*	*Readiness*	*Implementation*	*Impact*
East Asia & Pacific	40.06956	0.3238334	0.37370258	-0.060744257
Europe & Central Asia	41.40234	0.2690308	0.50195395	-0.088694891
Middle East & N.A.	42.24521	0.2343734	0.58306085	-0.106371030
North America	39.15587	0.3614031	0.28578021	-0.041582779
South Asia	37.34314	0.4359405	0.11134455	-0.003566914

Both *Readiness* and *Implementation* have consistently positive repercussions all throughout regardless of the region of the world studied. *Impact*, however, presents a negative effect in every instance. If we incorporate corruption levels as a control variable, *Implementation* coefficients become mostly negative, replicating the results from model 4.1. We must remember that corruption and income levels were the most highly correlated of the independent variables, lending credibility to the theory that *Implementation* and *Impact* have a negative effect on productivity for poorer countries because they generate new opportunities for graft which affect productivity negatively.

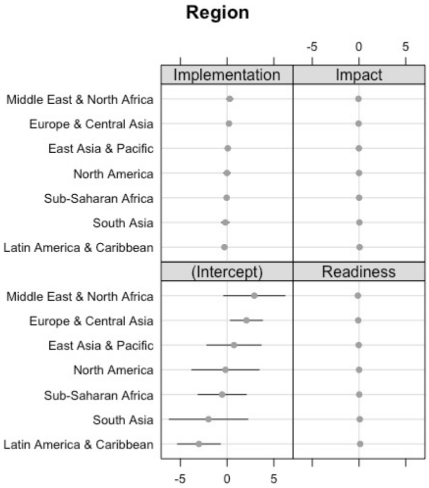


Figure 12 Random effects plot for each region

We incorporate corruption and group by income so we can also see how the behavior of corruption changes per income level instead of just seeing the sub-indexes.

$$Productivity = Implementation + Impact + Corruption + (Implementation + Impact + Corruption \mid Income) + \varepsilon \quad \textbf{Model 4.3}$$

Random effects

Groups	*Name*	*Variance*	*Std.Dev*	*Corr*
Income.2015	(Intercept)	1.481e+02	12.16780	
	Implementation	2.406e-03	0.04905	0.99

	Impact	3.379e-01	0.58128	-0.91	-0.97
Residual	4.657e+01	6.82421			

Fixed effects

	Estimate	*Std. Error*	*df*	*t value*	*Pr(>|t|)*
(Intercept)	23.08150	6.57265	3.26700	3.512	0.0342
Implementation	-0.17617	0.08923	6.81200	-1.974	0.0900
Impact	0.61338	0.30978	2.23900	1.980	0.1724
Corruption	0.75206	0.06237	15.87600	12.059	2.1e-09

Correlation of Fixed Effects

	(Intr)	Implmn	Impact
Implementtn	0.182		
Impact	-0.839	-0.499	
Corruption	-0.271	-0.325	-0.022

Random effect coefficients	*(Intercept)*	*Implementation*	*Impact*	*Corruption*
Low-income	18.552552	-0.1937277	0.8227503	0.7194574
Lower-middle-income	8.653815	-0.2349021	1.3089795	0.6313599
Upper-middle-income	29.728923	-0.1471867	0.2732281	1.2201910
High-income	35.390724	-0.1288585	0.0485794	0.5011723

It is revealing that when we incorporate *Corruption* into the fixed effects part of the model, *Readiness* loses all its significance, while *Implementation* and *Impact* become statistically significant (for α = 0.1). This would seem to suggest that once corruption levels are incorporated into our analysis, the level of preparedness for governments to incorporate data into their policies and procedures becomes irrelevant, while the sheer impact of those same policies and procedures in society are the only thing that actually have the power to increase efficiency, more strongly so at poorer countries, where perhaps the marginal improvements to be reaped are greater. This said, because *Corruption* and *Income* share a high correlation, these results should be taken with care.

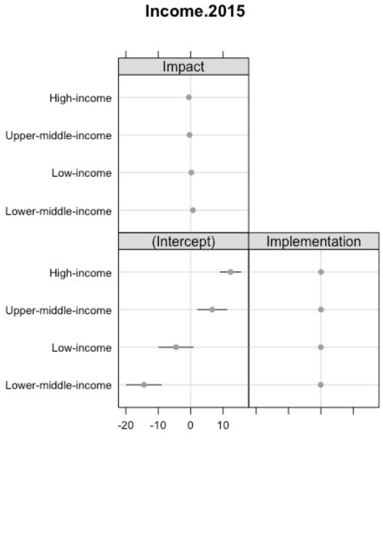


Figure 13 Random effects plot with corruption as control

Random Forests - Variable importance

So far we have been forcing the belonging of each country into groups according to their income level, region, corruption index or even HDI, but none of those classifications have been entirely satisfying due to lack of significance, collinearity or index weights. If we allow for hierarchical clustering of the sub-index scores (through R package *hclust*) and we draw a dendogram we can spot anywhere from 3 to 6 clusters (for illustration here we settle on 4) whose membership will be different from our previous classification.

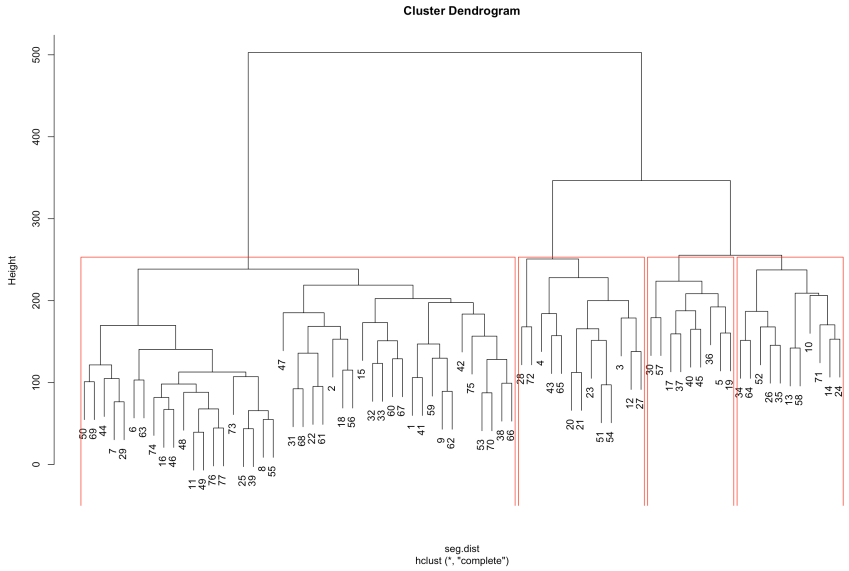


Figure 14 Clustering dendogram. The cophenetic correlation to the distance matrix is 0.7831

To have an idea of the relative importance of different attributes for the changes of productivity in a given country without first having segmented those countries, we run a regression Random Forest (through R package *randomForest*) which, after 10,000 iterations, gives us the following variable importance plot:

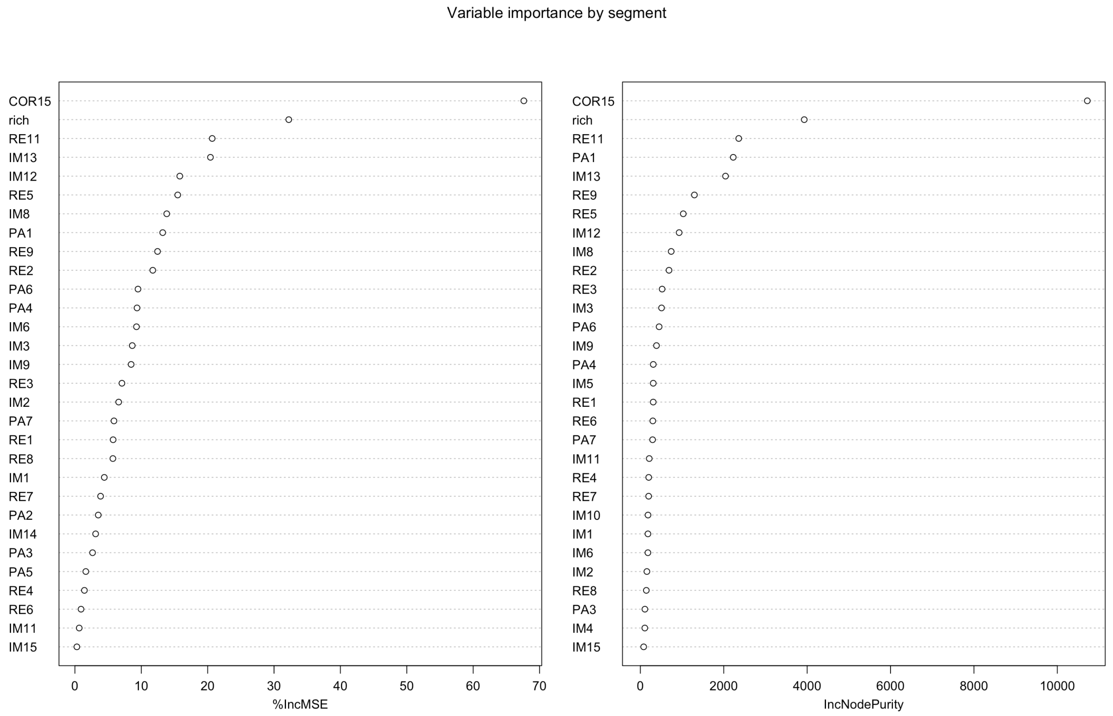


Figure 15 Random forest variable importance hierarchy

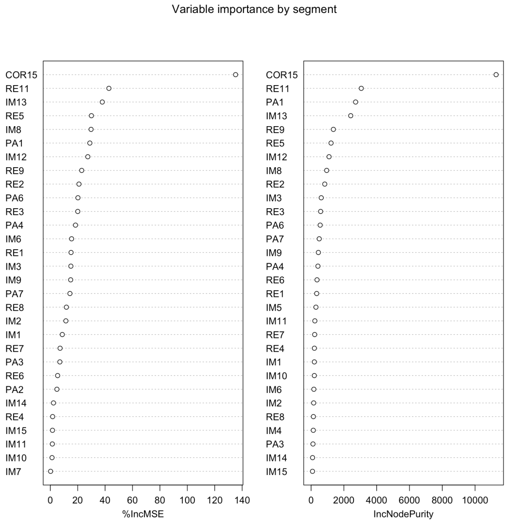


Figure 16 Importance without Income

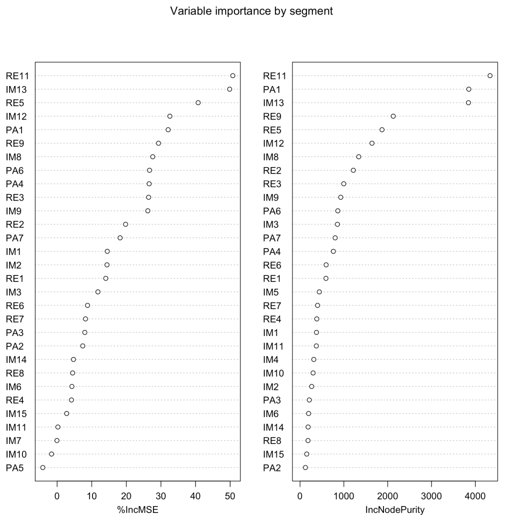


Figure 17 Without corruption or income

We guide ourselves by the predictive measures of the left column because node purity from the right column will be biased (Altmann, Toloşi, Sander, & Lengauer, 2010). Besides *Corruption* and *Income* (here a dummy variable *rich* for simplicity) some interesting results come up.

- *Readiness:* the most important influencer in government productivity according to our regression random forest is RE11 ("*Government Online Services Index*)". This is one of four components of the United Nations E-Government survey (United Nations Department of Economic and Social Affairs, 2016). It measures the quantity and quality of government services which are available online, from the perspective of a user/citizen. According to the UN report accompanying the survey, although this metric is positively correlated with income, this relationship has been becoming weaker in recent years as governments of all income levels provide more and more services online.

  Allowing certain bureaucratic procedures like tax filings to be conducted online has the advantage of helping combat fraud or increasing operational efficiencies to the tune of €7-30 billion and €120-200 billion per annum respectively according to McKinsey & Company (Manyika et al., 2011). Furthermore, besides operational benefits that repercuss chiefly back into governments, in the majority of cases the realization of certain procedures online will allow for the collection of data in a way that can benefit citizens two-fold: on the one hand, it will make procedures easier for them in the future thanks to improvements such as personalized and pre-populated government forms. On the other hand, it will help governments craft policies which will better cater to citizen's real needs as manifested though previously inaccessible patterns in their data. Of all *Readiness* metrics, it is the one that perhaps carries the most immediate and tangible benefits for citizens. All of this might help explain the preponderance of RE11 in the results.

  By contrast, RE6 ("*To what extent are civil society and information technology professionals engaging with the government regarding open data*?") and RE4 ("*To what extent does the country have a functioning right-to-information law?*")

have the least impact within *Readiness* and amongst the least impact overall.

RE6 aims to measure the extent to which government recognizes the need to involve civil society organizations, professionals and citizens in decisions regarding which data to collect and publish. While engaging civil society in government decisions is certainly a healthy thing for public administrators to do, the benefit of doing so may be overstated, and in any case falls mostly outside of what our *Government Effectiveness* index counts. It would be interesting to further segregate by income or corruption levels and compare whether RE6 holds more importance in less developed or more corrupt countries versus others which may be relatively better governed and so might see less benefit in "crowdsourcing" decisions or seeking consensus.

In RE4, right of information scores gauge whether information requests are answered in a timely manner, at reasonable costs, whether a justification is provided when requests are not complied with or, indeed whether they are complied with at all. In some ways this score is similar to RE6 and may behave similarly when examined within different country groups.

Lastly, the second most important *Readiness* variable is RE5, ("*To what extent is there a well-resourced open government data initiative in this country?*"*).* The definition is as follows: "An open data initiative is a plan by the government to release government data online to the public. It has four main features: (1) The government discloses data or information without request from citizens. This may be according to a release schedule or ad hoc; (2) The Internet is the primary means of disclosure. Mobile phone applications may also be used for disclosure; (3) Data is free to access and re-use, e.g. open licenses; (4) Data is in a machine-readable format to enable computer-based reuse, e.g. spreadsheet formats, Application Programming Interface (API).” RE5 contrasts in its importance to RE6 and RE4 in which while all assess whether government discloses its data and whether it does so in a reasonably affordable and accessible manner, RE5 evaluates so on the account of the government’s own initiative.

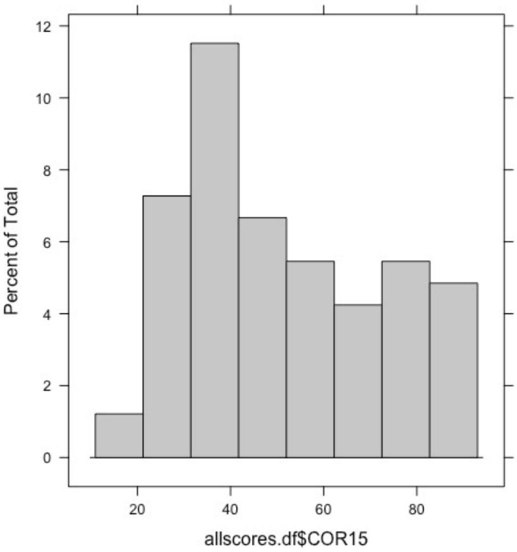


Figure 18 Corruption scores distribution, all countries

The distribution of *Corruption* scores in our sample is skewed towards the most corrupt countries, accepting the inherent imperfection in our data, this would suggest that if public sector productivity indeed benefits from administrations releasing their data, government corruption is not the key factor here, as the skewed distribution should elevate the relative importance of RE4 and RE6 and lower than of RE5.

Finally, the *Government Effectiveness* index addresses “Quality of bureaucracy / institutional effectiveness”, “Excessive bureaucracy / red tape”, “Bureaucratic quality” and “Policy instability”. As explained above, all of these would be at least partially affected positively both by RE11 and RE5 and so their relative importance appears easy to justify.

- *Implementation*: IM13 (“*Crime statistics”*) is the most important variable from this group. We already mentioned the study utilizing Los Angeles County Police Department and how analyzing the clustering of criminal activities allowed for the prediction of new, related happenings, and so permitted a more efficient re-allocation of police resources (Mohler et al., 2011). While it is easy to understand how accessible bulk crime data could improve citizen well-being, we did not consider security as part government output due to the difficulty of quantifying it, and in fact the World Governance Indicators distributes safety-related measurement across their "Political Stability and Absence of Violence/Terrorism" and "Rule of Law" sections, none of which we have included in our computations. The most immediate conclusion is that safety has a spillover effect in citizens' evaluation of other areas of government performance (that is, for equal performance in educational output, the more secure country would receive higher scores from its citizens in the educational area, not only in the safety area) or alternatively, crime statistics have latent benefits in that their usage helps the designing and carrying out of better policies in education, healthcare, bureaucracy and infrastructure.

IM15 ("*National election results"*) and IM11 ("*Health sector performance*") are the worst performing of the implementation variables. It seems apparent while having bulk, machine-readable election results does not carry much more benefit to public sector efficiency versus any other form of publication, such as press releases or reports. IM11 deserves more detailed consideration, in that healthcare has been hailed as one of the sectors that hold most promise to benefit from big data and advanced analytics (Henke et al., 2016) also (Desouza & Jacob, 2014) , (Jee & Kim, 2013), (European Big Data Value Association, 2014) and (Brown, Chui, & Manyika, 2011). Some of the areas highlighted by the abovementioned studies for healthcare to benefit from big data are comparative effectiveness research, clinical decision support systems, trial design or disease pattern recognition. All of the call for the exploiting of anonymized patient data to reap the benefits they promise. These data will be held by healthcare operators, not system administrators as the government is at their national, regional and even local levels. While public hospitals are funded and normally also operated by the government, they are the ultimate depositories of patient data, and so they are the ones who should collaborate with their data with pharmaceutical companies, universities and other professionals to extract the benefits of data in healthcare. Government involvement in the publication of healthcare performance data as defined by IM11 will reveal more about the managerial and bureaucratic soundness of public health provisioning than about disease patterns and optimal patient treatments.

- *Impact*: Our previous models already hinted at the comparable lack of relevancy that *Impact* metrics hold in explaining government efficiency variance. This seems supported by the fact that the highest-ranking metric, PA1 ("*Government e-participation index*") is only to be found well into the first third of important variables. This metric contains countries' levels of adaptation to new tools of e-participation and e-consultation. While it is correlated positively with income, it is less strongly so than the *Government Online Services Index,* with Lower-Middle-Income countries having adopted new online participatory tools faster than expected according to the UN report. While citizens can certainly have an impact in shaping policy, perhaps the impact is more indirectly felt than when consultative proceedings happen directly within the government thanks to greater access to data, simply because decisions will be faster to adopt and implement in the latter case than in the former.

Finally, *PA5* ("*To what extent has open data had a noticeable impact on increasing the inclusion of marginalised groups in policy making and accessing government services?*") is the metric within *Impact* that carries the least weight. This metric is not directly measured within the *Government Effectiveness* index although it is perhaps the section within the UN E-Government Survey which best captures the advancement of economic equality. The non-representation of marginalized groups can lead to "gross economic disadvantage" in the long term

(Young, 2002) but the effect of short term fluctuations are not as clear, so a higher degree of inclusion for marginalized groups might not carry as much direct impact in our index as other metrics.

Finite Mixture Models - Model 5

We introduce finite mixture models into our analysis under the presumption that latent class regressions will be helpful in our case where we have data from different sources that may have resulted in combinations of unobserved ("latent") distributions (Shaliza, 2012). Productivity, hence, could be explained as a convex combination, a mixture, of K component distributions, $f_1, f_2 \ldots f_K$

$$f(x) = \sum_{k=1}^{K} \lambda_k f_k(\mathrm{x}) \quad (8)$$

With the λ_k being the mixing weights, $\lambda_k > 0$, $\sum_k \lambda_k = 1$. In theory K could be equal to 1 and we would simply be estimating parameters. Alternatively, K could be equal to the number of observations (basically akin to traditional kernel density estimation). In practice, we will have something in between, and so we will choose a model that has enough components to fit the latent distributions accurately, but not so many that they cannot be estimated precisely and they have too much variance. With the help of package *flexmix* in R we iterate until the Expectation Maximization algorithm, through log-likelihood maximization, falls within a threshold pre-specified to equate ε (Leisch, 2004).

stepFlexmix(sP15 ~ RE1 + RE2 + RE3 + RE4 + RE5 + RE6 + RE7 + RE8 + RE9 + RE11 + IM1 + IM2 + IM3 + IM4 + IM5 + IM6 + IM7 + IM8 + IM9 + IM10 + IM12 + IM13 + IM14 + IM15 + PA1 + PA2 + PA3 + PA4 + PA5 + PA6 + PA7 + COR15, data = clusters4.df, control = list(verbose = 0), k = 1:35, nrep = 10) **Model 5.1**

Because log-likelihood maximisation by itself would simply lead us to choosing a number of components equal to the number of observations we balance this search for fit with a "penalized-likelihood" approach (Lampinen, Laaksonen, & Oja, 1997), such as the Bayesian Information Criterion (BIC) which seeks to minimise a composite score of the number of parameters and the number of observations. We prize BIC in our selection over Akaike's "An Information Criterion" (AIC) because the former puts more weight on the number of factors and thus, parsimony.

We let R determine the optimal level of components by attempting convergence at increasingly large numbers of factors, with a maximum of 35 (which is equal to the number of variables which compose the sub-indexes, plus the control variables). The optimal level of components is 25, with a BIC which is statistically significantly larger (Raftery, 1995) than the second best, 24 components. 25 components have a BIC score of 325 versus 420 for 24 components).

We have chosen not to group countries by income to avoid collinearity with corruption levels. Still, the program has found 2 clusters of sizes 41 and 36 respectively after 42 iterations, similar to when we grouped based on whether they were "rich" or not.

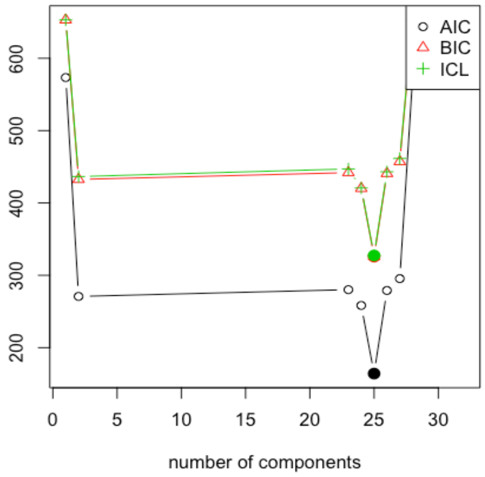


Figure 19 BIC, AID and ICL scores per number of components in finite mixture model

Having made sure that our observable variables are randomly distributed like normals or at least bell-shaped, we tend to assume that their sum within the latent distributions will also be normal, and so we specify a double Gaussian mixture distribution with the following predictors:

flexmix(sP15 ~ RE1 + RE2 + RE3 + RE5 + RE6 + RE7 + RE9 + RE11 + IM1 + IM2 + IM3 + IM4 + IM5 + IM8 + IM9 + IM12 + IM13 + PA1 + PA3 + PA4 + PA6 + PA7 + COR15, data = clusters4.df, k = 2,

model = list(FLXMRglm(sP15 ~ RE1 + RE2 + RE3 + RE5 + RE6 + RE7 + RE9 + RE11 + IM1 + IM2 + IM3 + IM4 + IM5 + IM8 + IM9 + IM12 + IM13 + PA1 + PA3 + PA4 + PA6 + PA7 + COR15, family= "gaussian"),
FLXMRglm(sP15 ~ RE1 + RE2 + RE3 + RE5 + RE6 + RE7 + RE9 + RE11 + IM1 + IM2 +IM3 + IM4 + IM5 + IM8 + IM9 + IM12 + IM13 + PA1 + PA3 + PA4 + PA6 + PA7 + COR15, family = "gaussian"))) **Model 5.2**

We converge at two clusters of 38 and 39 countries each after 19 iterations. These mixture models result in the following plot of countries against their productivity:

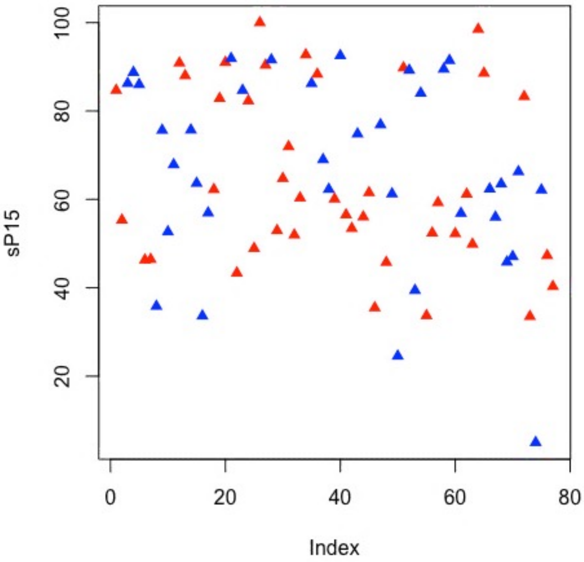


Figure 20 Productivity per predictor cluster

Fig. 19 does not seem to reveal clearly defined clusters. This might be either because the lack of depth of some variables means that their distribution is not perfectly normal, or because there are not enough instances of countries to guarantee identifying patterns in the variables. We get the predicted values of sP15 by selecting the component value of the cluster to which each data point belongs. The actual versus predicted results show the fit on the following regressed line against actual data points for the first Gaussian model:

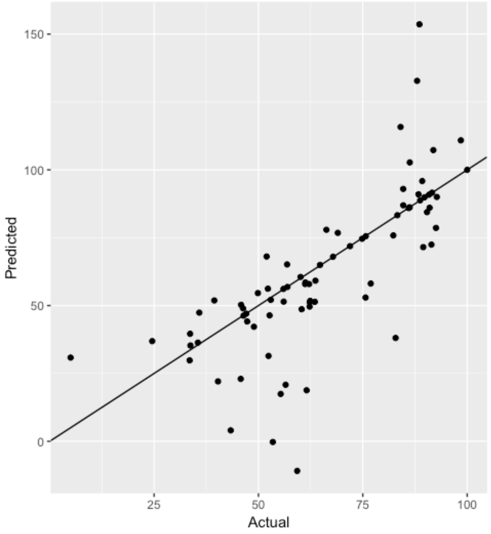


Figure 21 The root mean square error is 19.63152

There seem to be many outliers, adding to the difficulty of identifying latent patterns.

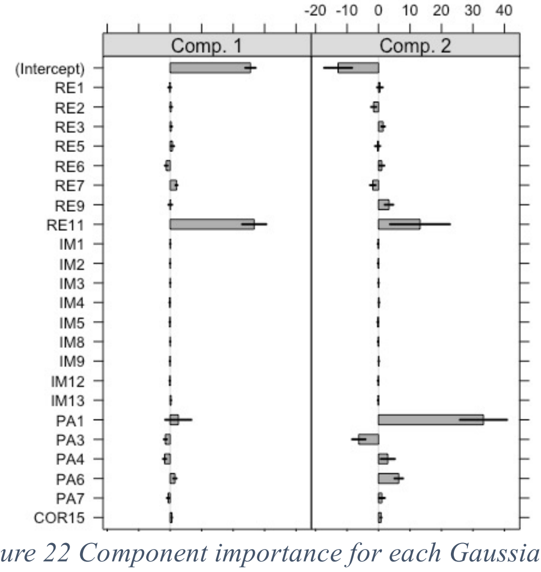
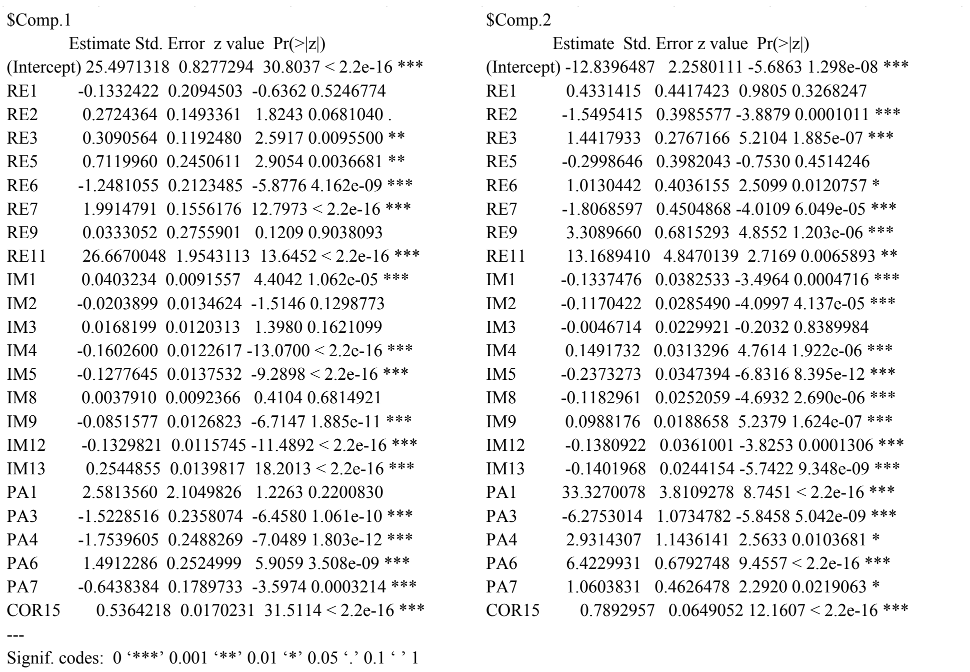


Figure 22 Component importance for each Gaussian model

The results are consistent with what we had seen in the results of the Random Forest, albeit with different weights since here the components are measured in two distinct models. This seems to confirm the outsize importance of RE11, PA1, RE9, PA4 and PA6, as well as diminish the importance of COR15, since the clustering into two groups takes away much of the importance of that metric.



```
$Comp.1
          Estimate Std. Error  z value  Pr(>|z|)
(Intercept) 25.4971318  0.8277294  30.8037 < 2.2e-16 ***
RE1        -0.1332422  0.2094503  -0.6362 0.5246774
RE2         0.2724364  0.1493361   1.8243 0.0681040 .
RE3         0.3090564  0.1192480   2.5917 0.0095500 **
RE5         0.7119960  0.2450611   2.9054 0.0036681 **
RE6        -1.2481055  0.2123485  -5.8776 4.162e-09 ***
RE7         1.9914791  0.1556176  12.7973 < 2.2e-16 ***
RE9         0.0333052  0.2755901   0.1209 0.9038093
RE11        26.6670048  1.9543113  13.6452 < 2.2e-16 ***
IM1         0.0403234  0.0091557   4.4042 1.062e-05 ***
IM2        -0.0203899  0.0134624  -1.5146 0.1298773
IM3         0.0168199  0.0120313   1.3980 0.1621099
IM4        -0.1602600  0.0122617 -13.0700 < 2.2e-16 ***
IM5        -0.1277645  0.0137532  -9.2898 < 2.2e-16 ***
IM8         0.0037910  0.0092366   0.4104 0.6814921
IM9        -0.0851577  0.0126823  -6.7147 1.885e-11 ***
IM12        -0.1329821  0.0115745 -11.4892 < 2.2e-16 ***
IM13         0.2544855  0.0139817  18.2013 < 2.2e-16 ***
PA1         2.5813560  2.1049826   1.2263 0.2200830
PA3        -1.5228516  0.2358074  -6.4580 1.061e-10 ***
PA4        -1.7539605  0.2488269  -7.0489 1.803e-12 ***
PA6         1.4912286  0.2524999   5.9059 3.508e-09 ***
PA7        -0.6438384  0.1789733  -3.5974 0.0003214 ***
COR15        0.5364218  0.0170231  31.5114 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
$Comp.2
           Estimate  Std. Error z value  Pr(>|z|)
(Intercept) -12.8396487   2.2580111 -5.6863 1.298e-08 ***
RE1          0.4331415   0.4417423  0.9805 0.3268247
RE2         -1.5495415   0.3985577 -3.8879 0.0001011 ***
RE3          1.4417933   0.2767166  5.2104 1.885e-07 ***
RE5         -0.2998646   0.3982043 -0.7530 0.4514246
RE6          1.0130442   0.4036155  2.5099 0.0120757 *
RE7         -1.8068597   0.4504868 -4.0109 6.049e-05 ***
RE9          3.3089660   0.6815293  4.8552 1.203e-06 ***
RE11         13.1689410   4.8470139  2.7169 0.0065893 **
IM1         -0.1337476   0.0382533 -3.4964 0.0004716 ***
IM2         -0.1170422   0.0285490 -4.0997 4.137e-05 ***
IM3         -0.0046714   0.0229921 -0.2032 0.8389984
IM4          0.1491732   0.0313296  4.7614 1.922e-06 ***
IM5         -0.2373273   0.0347394 -6.8316 8.395e-12 ***
IM8         -0.1182961   0.0252059 -4.6932 2.690e-06 ***
IM9          0.0988176   0.0188658  5.2379 1.624e-07 ***
IM12         -0.1380922   0.0361001 -3.8253 0.0001306 ***
IM13         -0.1401968   0.0244154 -5.7422 9.348e-09 ***
PA1         33.3270078   3.8109278  8.7451 < 2.2e-16 ***
PA3         -6.2753014   1.0734782 -5.8458 5.042e-09 ***
PA4          2.9314307   1.1436141  2.5633 0.0103681 *
PA6          6.4229931   0.6792748  9.4557 < 2.2e-16 ***
PA7          1.0603831   0.4626478  2.2920 0.0219063 *
COR15          0.7892957   0.0649052 12.1607 < 2.2e-16 ***
```

Figure 23 The majority of model 5.1 components are significant

We generate a density function to visualise the mixture models as well as a joint histogram featuring both models:

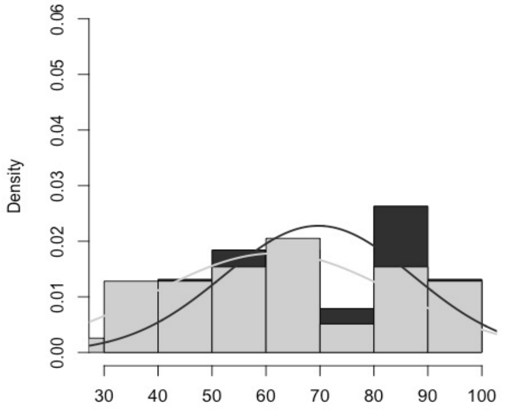


Figure 24 Density function of double Gaussian

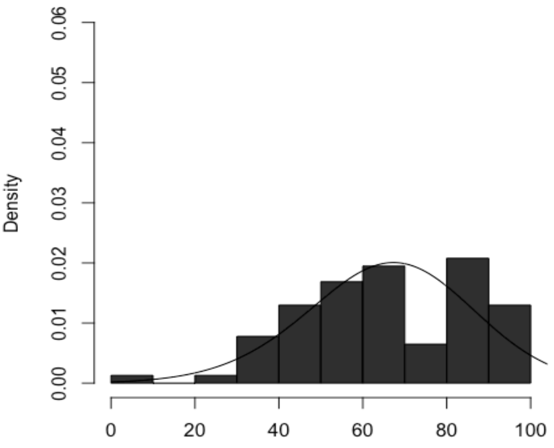


Figure 25 Joint density function of double Gaussian

It is normal for both functions to overlap to a certain extent. However, here they overlap excessively, and a problem is revealed whereby we lack enough data on the countries with a productivity score of 70 to 80.

Despite the lack of sufficient inputs, because two clusters seem to be an inappropriate form of grouping, we proceed with a triple Gaussian model to attempt to separate into more clearly defined clusters. It converges into 3 clusters of 23, 31 and 23 countries respectively after 34 iterations:

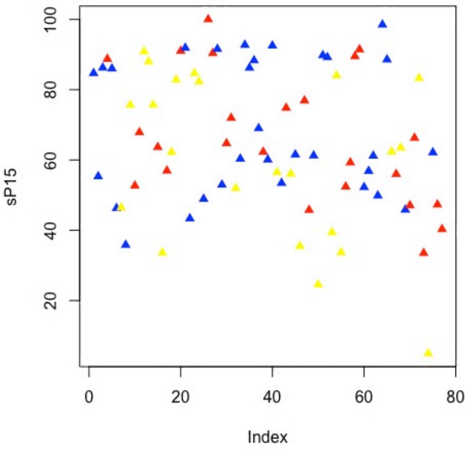


Figure 26 Productivity per triple predictor cluster

Once again, the triple Gaussian has not arrived at clearly visible separations. Unfortunately, our data prevents us from attempting a higher number of latent distributions, or a different type of them other than Gaussian. The difficulty of clustering countries into different composing groups might indeed be due to a lack of data points, but it might also hint at the real difficulty of assigning groups with such disparate variables at the early stage of a technology, when the development of one technological area does not necessarily mean the immediate adoption of a complementary part of the technology (Wozniak, 1993) and (Wozniak, 2012).

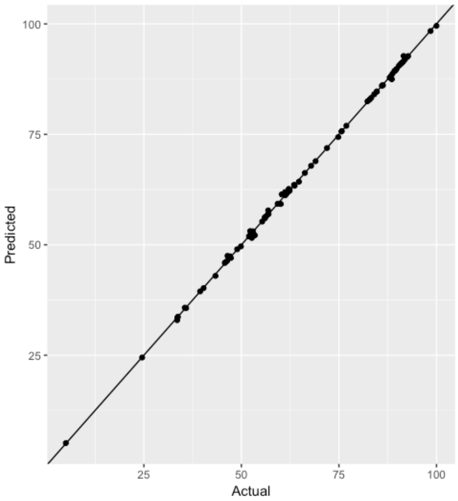


Figure 27 The root mean square error is 0.4076435

This almost perfect fit is more likely due to over-specifying components than to a great latent class match with the observable variables.

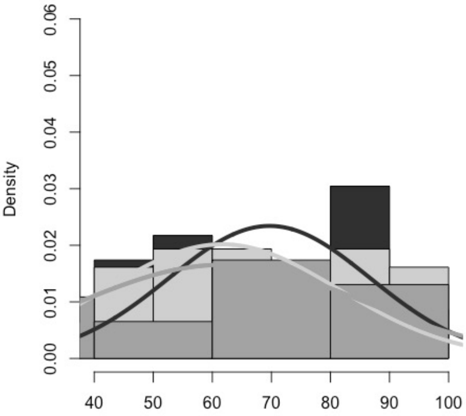


Figure 28 Density function of triple Gaussian

While the distributions approach normal shapes, we cannot glean the specificities of each group due to the lack of data which makes it hard to tell the clusters apart.

We must thus discard the finite mixture model approach for now, but this technique looks promising to be used in upcoming years when the amount of data collected for our predictors allow for further experimentation through the inclusion of more countries or years, and when countries' adoption of the technology matures uncovering more consistent patterns.

Principal Component Analysis

Principal Components Analysis (PCA) holds less interest for us because, although it can help us simplify the large number of components making up the sub-indexes, it is less helpful to understand their behavior under several scenarios due to many of the variables being partly correlated, unlike in a traditional PCA study, thus making the results less prescriptive and less interesting for policy making. Regardless, we proceed with the analysis to attempt to uncover associations, which we assess visually with factor loadings.

We use up all independent variables, plus *Corruption* for control. This time, we use the non-standardized scores and simply command R to center and scale everything. Through *prcomp* we arrive at a number of principal components whose role in explaining variance is illustrated in the next graph.

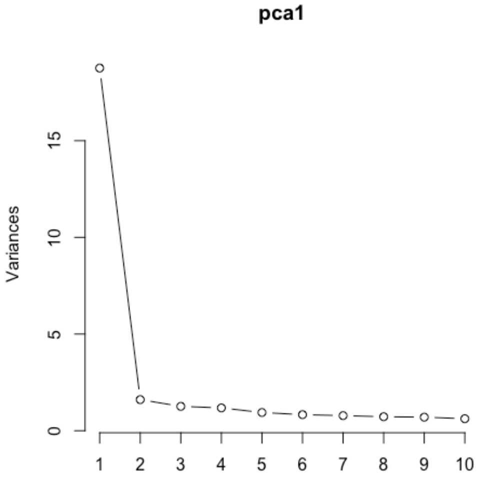


Figure 29 Variance explained per component

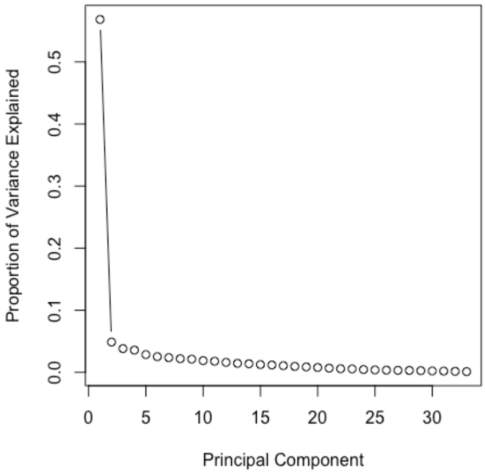


Figure 30 Proportion of variance explained per principal component

The first principal component already explains almost 60% of total variance, while the following 9 explain a further 15% of it. Visualizing the two main principal components:

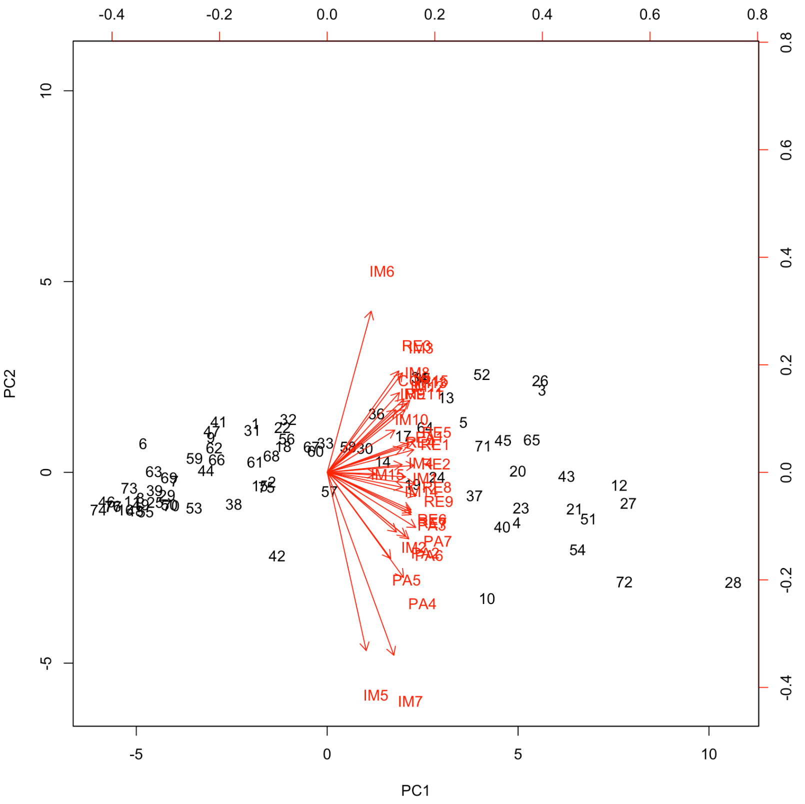


Figure 31 Principal components

In the rotated component matrix we see that RE9, PA7, RE8, RE5, RE2 and RE1 have the largest component loadings with respect to Component 1, in that order, while IM5, IM6, IM15, PA5 and IM7 have the smallest component loadings.

- **Component 1**: factor association.
  - *Strong association*: Existence of training for business and individuals, new data-based ventures being built by entrepreneurs, local governments running their own data initiatives, existence of well-resourced data initiatives, consistent data management and publication approach, existence of well-defined data policy.

- *Weak association*: detailed government budget (implementation), detailed data on government spend (implementation), national election results (implementation), inclusion of marginalized groups, company register (implementation).

Taken all together Component 1 could be called *Government Micromanaging*, since one could argue that the variables most strongly associated with it all represent proactive public sector involvement in their civil and business communities. Conversely, it is most weakly associated with variables that represent the mere existence and accessibility of data, perhaps in the hope that the economic and social impact materializes organically with a hands-off, *laissez-faire* approach.

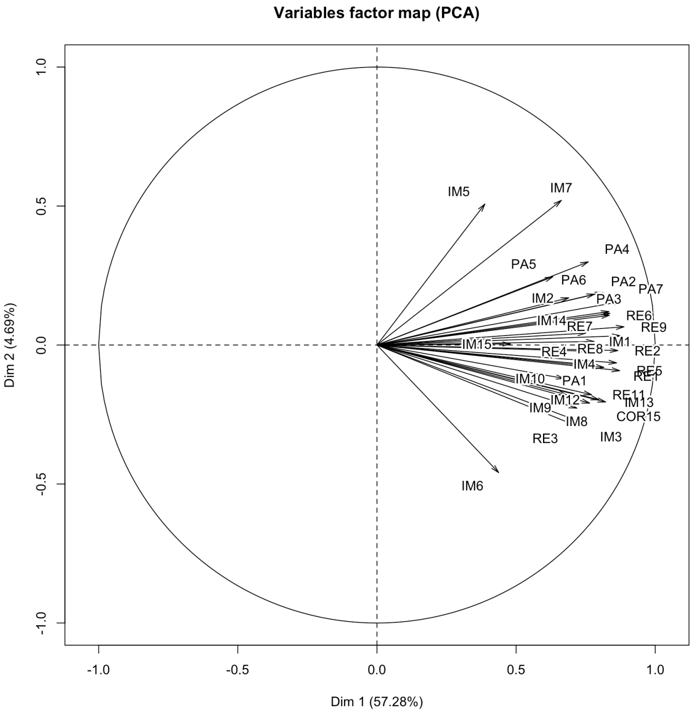
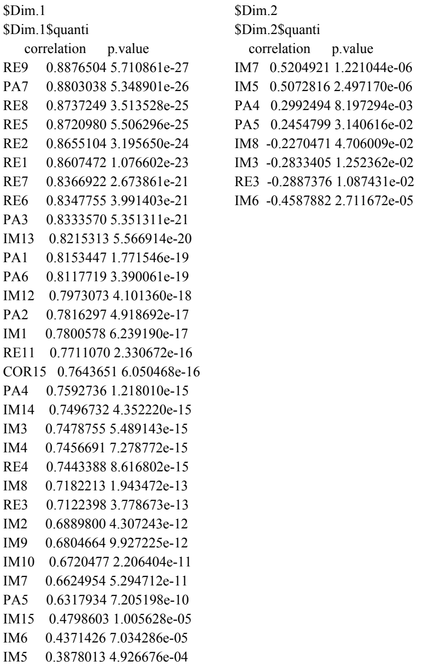


Figure 32 Alternative visualisation of components with FactoMineR

In the rotated component matrix we see that IM7, IM5, and PA4 have the largest component loadings with respect to Component 2, in that order, while IM6, RE3, and IM3 have the smallest component loadings.

- **Component 2**: factor association.
    - *Strong association:* company register (implementation), detailed government budget (implementation), environmental impact.
    - *Weak association*: detailed data on government spend (implementation), robust regulatory framework for data protection, detailed census data (implementation).

Taken all together, Component 2 could be called *Civil Impact* since the variables most strongly correlated with it seem to represent opportunities for civil society to act on government data (be them business making decisions based on company registration data or areas where governments plan to spend money) while it is most weakly correlated with variables that are most useful to create impact for government itself rather than civil society (as in *past* spending statistics rather than *planned* spending, or census data).



```
$Dim.1
$Dim.1$quanti
     correlation     p.value
RE9     0.8876504 5.710861e-27
PA7     0.8803038 5.348901e-26
RE8     0.8737249 3.513528e-25
RE5     0.8720980 5.506296e-25
RE2     0.8655104 3.195650e-24
RE1     0.8607472 1.076602e-23
RE7     0.8366922 2.673861e-21
RE6     0.8347755 3.991403e-21
PA3     0.8333570 5.351311e-21
IM13    0.8215313 5.566914e-20
PA1     0.8153447 1.771546e-19
PA6     0.8117719 3.390061e-19
IM12    0.7973073 4.101360e-18
PA2     0.7816297 4.918692e-17
IM1     0.7800578 6.239190e-17
RE11    0.7711070 2.330672e-16
COR15   0.7643651 6.050468e-16
PA4     0.7592736 1.218010e-15
IM14    0.7496732 4.352220e-15
IM3     0.7478755 5.489143e-15
IM4     0.7456691 7.278772e-15
RE4     0.7443388 8.616802e-15
IM8     0.7182213 1.943472e-13
RE3     0.7122398 3.778673e-13
IM2     0.6889800 4.307243e-12
IM9     0.6804664 9.927225e-12
IM10    0.6720477 2.206404e-11
IM7     0.6624954 5.294712e-11
PA5     0.6317934 7.205198e-10
IM15    0.4798603 1.005628e-05
IM6     0.4371426 7.034286e-05
IM5     0.3878013 4.926676e-04
```

```
$Dim.2
$Dim.2$quanti
   correlation     p.value
IM7   0.5204921 1.221044e-06
IM5   0.5072816 2.497170e-06
PA4   0.2992494 8.197294e-03
PA5   0.2454799 3.140616e-02
IM8  -0.2270471 4.706009e-02
IM3  -0.2833405 1.252362e-02
RE3  -0.2887376 1.087431e-02
IM6  -0.4587882 2.711672e-05
```

Figure 33 Variable correlation to components

Finally, we look where countries are situated within those two axes, *Government Micromanaging* (“*dim1*”) and *Civil Impact* (“*dim2*”).

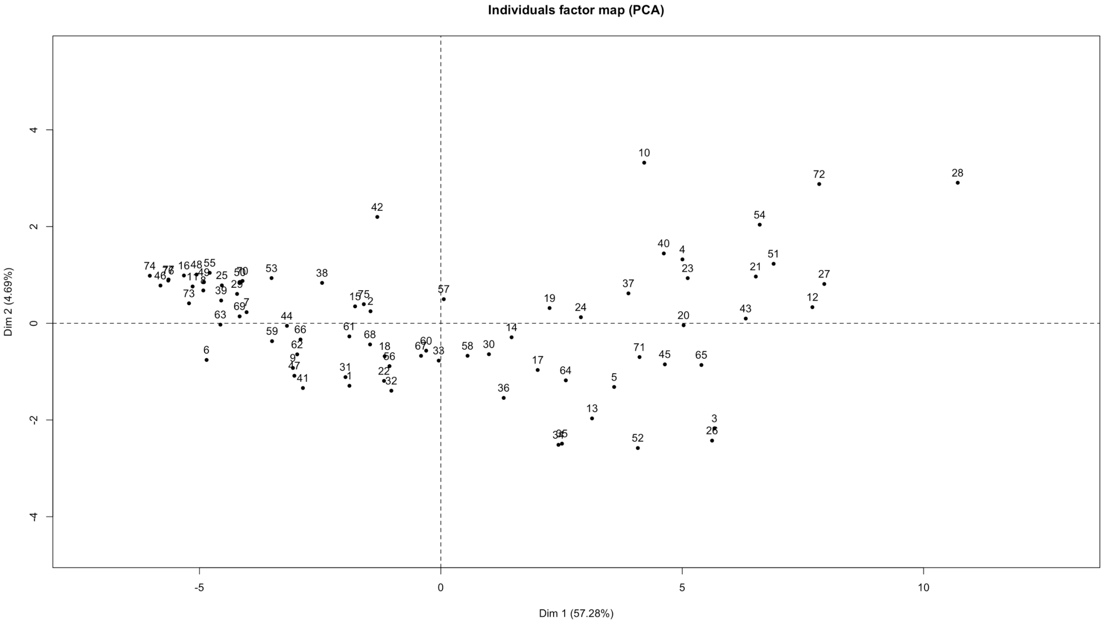


Figure 34 Countries mapped against their PCA components score

Component	Performance	Country	Productivity
1. Micromanaging	Good	United Kingdom	91.59
		France	90.38
		United States	83.28
		Canada	90.87
		Netherlands	89.78
		New Zealand	84.05
	Bad	Yemen	0
		Mali	35.48
		Zambia	47.31
		Zimbabwe	40.31
		Cameroon	33.63
		Venezuela	33.52

Component	Performance	Country	Productivity
2. Civil impact	Good	Brazil	52.67
		United Kingdom	91.59
		United States	83.28
		Kenya	54.44
		New Zealand	84.05
		Japan	92.53
	Bad	Norway	89.23
		Ireland	92.73
		Iceland	86.22
		Switzerland	88.00
		Israel	88.33
		Indonesia/Belgium	52/86

There seems to be a clear positive association between countries which micromanage their involvement locally with data initiatives, training, etc. and high productivity. The mere recollection of data does not excise positive impact in the productivity of the countries listed.

More interestingly, although the second component barely explains 5% of the total variance, its results are more mixed. High-productivity governments are found in the group with “good” civil impact scores together with mediocre performers such as Brazil and Kenya. Conversely, “Bad” civil impact governments are mostly high-productivity with the expectation of Indonesia, here in a tie in the component score with Belgium. This might suggest that “bad” civil impact (or in other words, “good” government impact) generally is associated with countries with high-productivity public sectors, while countries that have put an emphasis in making tools available for civil impact to happen regardless of their intervention have seen mixed results.

Both component score results taken together seem to suggest that active involvement in the development of data initiatives, particularly training, local initiatives and entrepreneurship, together with government-impact oriented policies, have a larger positive effect in productivity than passive measures such as making different types of data available in conveniently accessible ways.

Interaction effects - Model 6

To wrap up what we have learned from the previous models, we select the most important variables highlighted by Random Forest, Finite Mixture Models and Principal Component Analysis, and examine their behaviour together with the interaction effects of the two main control variables, *Corruption* and *Income*, both simplified with a dummy (For *Income*, *rich1* = High-income, else = *rich0*. In the case of *Corruption, corrupt0* ≥ $\text{median}_{\text{corruption}}$, else *corrupt1*).

Model 6 - 2015

	Dependent variable:
	Productivity 2015
Scores 2015	0.243^{***}
	(0.061)
IM12	0.144^{***}
	(0.050)
RE8	-0.102^{**}
	(0.048)
RE3	0.080^{**}
	(0.040)
PA2	-0.097^{*}
	(0.055)
RE9	0.154^{*}
	(0.082)
rich0:corrupt	-21.282^{***}
	(2.570)
rich1:corrupt	-16.862^{***}
	(4.386)

Constant	49.849***
	(4.108)
R^2	0.842
Adjusted R^2	0.824
Residual Std. Error	8.627 (df = 68)
F Statistic	45.402*** (df = 8; 68)
Note:	*p<0.1; **p<0.05; ***p<0.01

To arrive at a final model, we have only retained the most significant variables. We try our model in the datasets from 2014 and 2013, the only other years for which we have the independent variable data, as published in the first two editions of the ODB. We must exclude PA2 as that variable was sourced from the ODB and it was not studied in the first two editions. All variables have been standardised in the same way.

Model 6 - 2014

	Dependent variable:
	Productivity 2014
Scores 2014	0.264***
	(0.083)
IM12	0.035
	(0.047)
RE8	-0.053
	(0.060)
RE3	0.183***
	(0.061)
RE9	0.111
	(0.080)
rich0:corrupt	-16.990***
	(3.376)
rich1:corrupt	-23.992***
	(4.588)
Constant	38.565***
	(5.058)

R^2	0.826
Adjusted R^2	0.808
Residual Std. Error	9.464 (df = 69)
F Statistic	46.767^{***} (df = 7; 69)
Note:	*p<0.1; **p<0.05; ***p<0.01

Model 6 - 2013

	Dependent variable:
	Productivity 2013
Scores 2013	0.090^{*}
	(0.053)
IM12	0.116^{**}
	(0.049)
RE8	-0.006
	(0.053)
RE3	0.175^{***}
	(0.048)
RE9	0.057
	(0.072)
rich0:corrupt	-24.285^{***}
	(2.976)
rich1:corrupt	-27.605^{***}
	(4.899)
Constant	47.564^{***}
	(3.760)
R^2	0.810
Adjusted R^2	0.790
Residual Std. Error	10.078 (df = 69)
F Statistic	41.890^{***} (df = 7; 69)
Note:	*p<0.1; **p<0.05; ***p<0.01

Most of the indicators coincide with those highlighted by Random Forest and FMM, and they retain their significance in different years. All models feature strong R^2.

6. CONCLUSION

Even if Big Data is becoming old news in the business world, its introduction into the public sector is still very much at the early stages, let alone the introduction of its related disciplines, Machine Learning and Artificial Intelligence. Still, we have proven that governments do indeed benefit from integrating data into their operations, with a 1 point increase in their composite data score resulting in 0.16 increase in their productivity (according to model 1.2).

Model 3 shows that *Implementation* and *Impact* can have a negative impact that seems to be related to corruption because, in the case of corrupt governments, the availability of certain types data might aid graft. A further study could look into the added benefit of improving country scores beyond the "*government data exists*" stage in *Implementation* because there seems to be smaller marginal benefits to data being available "*in bulk"* or "*in machine readable formats*". Therefore, the investment needed to bring systems up to par may only pay off for countries in the later stages of their development.

Random Forests gave us the first insights into what low-level variables mattered most to productivity, with the importance of some of them being again vindicated with other methods in later stages. These most-relevant variables are the *Government Online Services Index,* the existence and availability of *crime datasets*, the existence of *well-resourced open government data initiatives*, the *Government E-participation Index*, and the *availability of training for individuals or businesses wishing to increase their skills or build businesses*.

Principal Component Analysis regrouped our variables into several components of which we retained the most important two. Once we assign a category to each of those components based on the variables they are most and least strongly correlated with, we position them into two axes and examine where countries fall in the resulting plane. The result suggests that high-productivity governments fall closer to the most positive values of axis 1, which we associate with governments *micromanaging* the development of these nascent data-based technologies both in the public administration and in public society, versus a more hands-off approach consisting on the publication of data and expecting for industries or civil impact to grow organically around it. Conversely, high-productivity governments also appear to put more attention into the development of data tools and policies for themselves than for civil society.

Finally, the difficulty of clustering latent factors with Finite Mixture Models reminds us that with this being a new technology, the many factors that point to its adoption do not advance at the same rate, in line with previous studies have suggested. Patterns will become more consistent and easier to identify with time, and the data integration variables that most affect productivity will also change as the technology matures. All being good reasons to revisit this study in a few years' time.

7. BIBLIOGRAPHY

1. Afonso, A., Schuknecht, L., & Tanzi, V. (2005). Public sector efficiency: An international comparison. *Public Choice*, *123*(3–4), 321–347. https://doi.org/10.1007/s11127-005-7165-2
2. Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/10.1093/bioinformatics/btq134
3. Antonio M. Conti, Stefano Neri, A. N. (2017). *Low inflation and monetary policy in the Euro Area*.
4. Atkinson, A. B. (2005). The Atkinson review: final report. Measurement of government output and productivity for the national accounts. https://doi.org/10.1787/9789264009011-en
5. Brown, B., Chui, M., & Manyika, J. (2011). Are you ready for the era of "big data"? *McKinsey Quarterly*, *4*(October), 24–35. https://doi.org/00475394
6. Caulier-grice, J., Mulgan, G., Open, T., Programme, I., Sdud, L., Dwrv, G. H., … Vrq, O. (2015). *Open Data Barometer Global Report - 3rd Edition*.
7. *Corruption Perceptions Index*. (2016). *Transparency International*. https://doi.org/978-3-943497-18-2
8. Decisions Regarding Monetary Policy Implementation - December 14, 2016. (n.d.). Federal Reserve.
9. Desouza, K. C., & Jacob, B. (2014). Big Data in the Public Sector: Lessons for Practitioners and Scholars. *Administration & Society*, (September), 95399714555751. https://doi.org/10.1177/0095399714555751
10. European Big Data Value Association. (2014). European Big Data Value Strategic Research and Innovation. *BigDataValu.eu*, (July), 1–45. Retrieved from http://www.nessi-europe.eu/Files/Private/EuropeanBigDataValuePartnership_SRIA__v099 v4.pdf
11. Figari, F., Paulus, A., Sutherland, H., Avram, S., Leventi, C., Levy, H., … Rastrigina, O. (2015). The design of fiscal consolidation measures in the European Union: Distributional effects and implications for macroeconomic recovery, (March).
12. Giavazzi, F., & Pagano, M. (1995). *Non-Keynesian Effects of Fiscal Policy Changes: International Evidence and the Swedish Experience*.
13. Goel, R. K., & Nelson, M. A. (1998). Corruption and Government Size: A Disaggregated Analysis. *Public Choice*, *97*(1), 107–20. https://doi.org/10.1023/A:1004900603583
14. Hale, T. (Financial T. (2017). European Central Bank ECB bond buying slide raises taper questions, pp. 1–6. Retrieved from https://www.ft.com/content/c192cb00-dccf-11e6-86ac-f253db7791c6
15. Head, J. G. (1970). Exchange Theorie of the Public Economy. *Public Finance Analysis*, *29*(1970), 112–121.

16. Henke, N., Bughin, J., Chui, M., Manyika, J., Saleh, T., Wiseman, B., & Sethupathy, G. (2016). The Age of Analytics: Competing in a Data-Driven World, (December), 136.

17. Jee, K., & Kim, G. H. (2013). Potentiality of big data in the medical sector: Focus on how to reshape the healthcare system. *Healthcare Informatics Research*, *19*(2), 79–85. https://doi.org/10.4258/hir.2013.19.2.79

18. Kaufmann, D., Kraay, A., & Mastruzzi, M. (2011). The Worldwide Governance Indicators: Methodology and Analytical Issues. *Hague Journal on the Rule of Law*, *3*(2), 220–246. https://doi.org/10.1017/S1876404511200046

19. Lampinen, J., Laaksonen, J., & Oja, E. (1997). Neural Network Systems, Techniques and Applications in Pattern Recognition. *Laboratory of Computational Engineering*.

20. Leisch, F. (2004). FlexMix: A General Framework for Finite Mixture Models and Latent Class Regression in R. *J. Stat. Softw.*, *11*(8), 1–18. https://doi.org/http://dx.doi.org/10.18637/jss.v011.i08

21. Long, J. S., & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, *54*(May), 217–224. https://doi.org/10.2307/2685594

22. Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung Byers, A. (2011). Big data: The next frontier for innovation, competition, and productivity. *McKinsey Global Institute*, (June), 156. https://doi.org/10.1080/01443610903114527

23. Martínez, E. A., Rubio, M. H., Martinez, R. M., Patane, D., Zerbe, A., Kirkpatrick, R., & Luengo-oroz, M. (2016). Measuring Economic Resilience to Natural Disasters with Big Economic Transaction Data. *Bloomberg Data for Good Exchange Conference*.

24. Mauro, P. (1995). CORRUPTION AND GROWTH. *Quarterly Journal of Economics*, (August).

25. Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-Exciting Point Process Modeling of Crime. *Journal of the American Statistical Association*, *106*(493), 100–108. https://doi.org/10.1198/jasa.2011.ap09546

26. Monetary Policy Decisions - January 19, 2017. (n.d.). European Central Bank.

27. Musgrave, R. A. (1939). The Voluntary Exchange Theory of Public Economy. *The Quarterly Journal of Economics*, *53*(2), 213–237. https://doi.org/10.1093/cje/bem005

28. Raftery, A. E. (1995). Bayesian Model Selection in Social Research. *Sociological Methodology*, *25*, 111–163.

29. Sala-I-Martin, X. X. (1997). I Just Ran Two Million Regressions. *The American Economic Review*, *87*(2), 178–183.

30. Santis, R. A. De. (2016). *Impact of the asset purchase programme on euro area government bond yields using market news*. *Technology*.

31. Schnabel, P., de Kam, F., Kuhry, B., & Pommer, E. (2004). *Public Sector*

Performance. An international comparison of education, health care, law and order and public administration. Encyclopedia of Social Measurement (Vol. 3). The Hague: Social and Cultural Planning Office. https://doi.org/10.1016/B0-12-369398-5/00475-8

32. Shaliza, C. (2012). Mixture Models. *Advanced Data Analysis (Lecture Notes)*, 390–419.
33. United Nations Department of Economic and Social Affairs. (2016). *UN E-government survey 2016. E-Government in Support of Sustainable Development.* https://doi.org/10.1016/S1369-7021(02)00629-6
34. Wozniak, G. D. (1993). Joint Information Acquisition and New Technology Adoption: Late Versus Early Adoption. *The Review of Economics and Statistics*, *75*(3), 438–445. https://doi.org/10.2307/2109457
35. Wozniak, G. D. (2012). Human Capital , Information , and the Early Adoption of New Technology. *Technology*, *22*(1), 101–112. https://doi.org/10.2307/145869
36. Young, I. M. (2002). Inclusion and Democracy. *Chicago Journals*, *112*(3), 646–650.
37. Mayer-Schönberger, V. a. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* Boston: Houghton Mifflin Harcourt.

论文独创性声明

本论文是我个人在导师指导下进行的研究工作所取得的研究成果。论文中除特别加以标注和致谢的地方外，不包含其他人或其他机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

作者签名： 日期：2017 年 3 月 30 日

论文使用授权声明

本人完全了解复旦大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。保密论文在解密后遵守此规定。

作者签名： 导师签名： 日期：2017 年 3 月 30 日

论文独创性声明

本论文是我个人在导师指导下进行的研究工作及取得的研究成果。论文中除了特别加以标注和致谢的地方外，不包含其他人或其它机构已经发表或撰写过的研究成果。其他同志对本研究的启发和所做的贡献均已在论文中作了明确的声明并表示了谢意。

作者签名：____________　日期：__________

论文使用授权声明

本人完全了解复旦大学有关保留、使用学位论文的规定，即：学校有权保留送交论文的复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容，可以采用影印、缩印或其它复制手段保存论文。保密的论文在解密后遵守此规定。

作者签名：___________　导师签名：____________　日期：___________