TALLINN UNIVERSITY OF TECHNOLOGY
Faculty of Information Technology

Kaisa Vaino 152892IABM

DATA COLLECTION AND PRE-ANALYSIS FOR RESEARCH PROJECT

Master's thesis

Supervisor: Tarmo Veskioja
Researcher
Department of Software Science

Co-supervisor: Aaro Hazak
Professor
Department of Economics and Finance

Tallinn 2018

TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Kaisa Vaino 152892IABM

ANDMETE KOGUMINE JA EELANALÜÜS UURIMISPROJEKTI JAOKS

Magistritöö

Juhendaja:	Tarmo Veskioja Teadur Tarkvarateaduse instituut
Kaasjuhendaja:	Aaro Hazak Professor Rahanduse ja majandusteooria instituut

Tallinn 2018

Author’s declaration of originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Kaisa Vaino

07.05.2018

Abstract

The main aim of this thesis is to design and implement source data repository for the research project, which investigates the role of institutional factors in knowledge-based economic development, and provide tools that facilitate data pre-analysis.

The main problems confronted in the thesis relate to the general problems such as database design and implementation as well as to the problems specific to the research project, such as the identification of data sources and selection of appropriate data.

The main outcome of the thesis is a database implemented on MS Access platform, filled with data relevant for the research. Database applications allow to generate sets of data in panel format that is suitable for further econometric modelling in data analysis software such as STATA, Eviews or R. Database applications which enable to obtain an overview of the selected indicators from various aspects and assess their suitability for further analysis were designed to complement the indicator pre-selection and facilitate data pre-analysis process. Pre-selected indicators and applications designed during this thesis help to define and/or potentially drive the next stages of the research project.

The thesis is written in English and contains 95 pages of text, 8 chapters, 14 figures and 4 tables. Database designed and implemented on MS Access platform forms inseparable part of this thesis.

Annotatsioon

Andmete kogumine ja eelanalüüs uurimisprojekti jaoks

Antud töö peamiseks eesmärgiks on disainida ja realiseerida andmebaas, mis koondab teadmuspõhise majandusarengu uurimisega seotud teadustöö jaoks vajalikke algandmeid ning luua rakendused andmete eelanalüüsi lihtsustamiseks.

Töös käsitletud probleemide seas on nii andmebaasi disaini ja realisatsiooniga seonduvaid küsimusi kui ka spetsiifilisi teadustöö teemaga seonduvaid küsimusi ning lahendusi.

Töö peamiseks väljundiks on MS Access platvormil realiseeritud andmebaas, mis koondab endas teadustöö jaoks relevantseid andmeid. Andmebaas võimaldab lihtsalt genereerida andmekogumeid paneelandmete formaadis, mis sobivad edasiseks ökonomeetriliseks modelleerimiseks valitud andmeanalüüsi platvormil, näiteks STATA, Eviews või R. Lisaks indikaatorite eelvalikule on andmete eelanalüüsi läbiviimise hõlbustamiseks realiseeritud ka vahendid, mis võimaldavad valitud indikaatoritest ülevate saada ning hinnata nende sobivust edasiseks analüüsiks eri aspektidest lähtuvalt. Antud töö käigus loodud vahendid ja läbi viidud andmete eelvalik aitavad defineerida teadustöö täpsemaid uurimissuundi ja viia läbi teadustöö järgmisi etappe.

Lõputöö on kirjutatud inglise keeles ning sisaldab teksti 95 leheküljel, 8 peatükki, 14 joonist, 4 tabelit. Lõputöö lahutamatuks osaks on töö käigus disainitud ja MS Access platvormil realiseeritud andmebaas.

List of abbreviations and terms

Abbreviations

ADB	Asian Development Bank
APEC	Asia-Pacific Economic Cooperation
API	Application Programming Interface
CRISP-DM	Cross Industry Standard Process for Data Mining
ETL	Extract Transform Load
GDP	Gross Domestic Product
ICT	Information and Communication Technologies
IMF	International Monetary Fund
KAM	Knowledge Assessment Methodology
KBD	Knowledge-Based Development
KBE	Knowledge-Based Economy
KEI	Knowledge Economy Index
MERIT	Maastricht Economic and Social Research Institute
ODBC	Open Database Connectivity
OECD	Organization for Economic Co-operation and Development
OLAP	On-Line Analytical Processing
OLTP	On-Line Transaction Processing
PISA	Programme for International Student Assessment
R&D	Research and Development
RDBMS	Relational Database Management Systems
SQL	Structured Query Language
VBA	Visual Basic Application
WDI	World Development Indicators
XML	Extensible Markup Language

Terms

Final output table	Refers to the database table *master_cross_table*, which holds selected data in panel data format
Research project	Research project investigating institutional factors of knowledge-based economic development
This thesis	This Master thesis

Table of contents

1 Introduction 11

1.1 Purpose 13

1.2 Methodology 14

1.3 Overview of the study 14

2 Methodology 16

2.1 General methodology for econometric modelling 16

2.2 Thesis methodology 17

3 Data needs and data sources 20

3.1 Background of the data needs 20

3.2 Hypotheses 24

3.3 Indicators for knowledge-based economy 25

3.3.1 World Bank Knowledge Assessment Methodology 25

3.3.2 OECD framework 27

3.3.3 Knowledge Economy Indicator project by MERIT 28

3.3.4 APEC framework 29

3.3.5 Critique of KBE indicators 30

3.4 Synthesis of KBE framework analysis 31

3.5 Identified data requirements 33

4 Designing and implementing database 35

4.1 Database type 37

4.2 Requirements for the database 38

4.3 Selection of the technical solution 39

4.4 Description of the technical solution and data flows 40

4.5 Database design process 42

4.6 Data model 43

4.7 ETL processes 47

5 Data pre-analysis 51

5.1 Feature selection 51

5.2 Design and implementation of pre-analysis tools 56

5.2.1 Indicator descriptive statistics 57

5.2.2 Indicator availability.. 58

5.2.3 Conformity analysis.. 59

6 Delivery 63

7 Conclusion.. 64

8 Kokkuvõte 66

References 68

Appendix 1 – Database base table descriptions.. 71

Appendix 2 – Modules 75

Appendix 3 – Queries.. 81

Appendix 4 – Instructions for the database 85

Appendix 5 – Overview of pre-selected indicators 90

List of figures

Figure 1 Phases of econometric modelling based on (Paas, 1995) and (Brooks, 2008) 17

Figure 2 Thesis process map 18

Figure 3 Variables of KAM... 26

Figure 4 World Bank KAM Basic Scorecard... 27

Figure 5 Solution architecture 41

Figure 6 Snapshot of the database 43

Figure 7 Data model 44

Figure 8 Rule violation messages.. 47

Figure 9 Feature selection process... 52

Figure 10 Pre-selection tables and queries for restoring the pre-selection..................... 56

Figure 11 Snapshot of view View1_selected_indicator_stats 57

Figure 12 Snapshot of view View2_indicator_period_coverage 58

Figure 13 Snapshot from initial unordered dataset (example with 6 countries and 7 indicators).. 60

Figure 14 Conformity analysis (example with 6 countries and 7 indicators)................. 60

List of tables

Table 1 Comparison of KBE measurement frameworks 32
Table 2 Sample indicators of KBE 32
Table 3 Overview of database tables 44
Table 4 Overview of indicators included in the database 53

1 Introduction

This thesis is part of a research project that seeks to investigate the economic development in select Southeast Asian transition economies towards the knowledge-based economy (KBE) and society. KBE refers to the use of ‘knowledge’ to produce economic benefits (Günther, 2005). The research project puts special emphasis on understanding the role of different institutions in knowledge intensive economic development. Ample research (Timmer (2006), North (1990), Acemolu et al (2005)) on developed countries suggests that institutional efficiency and economic regime are key determinants towards knowledge-based development (KBD). The research is concentrated on Cambodia, Laos and Vietnam - emerging market economies with shared history of wars, political instability and regime changes, which has left the countries far behind in economic development. The research project investigates the mechanisms, regulatory incentives and challenges in transferring knowledge into economic value and aims to identify and outline regulatory measures to address market frictions and inefficiencies on the path towards a knowledge economy (Hazak, forthcoming).

This thesis focuses on the critical phase of the research project: data collection and data pre-analysis. Measurement of KBE and knowledge intensity is a complex topic – no universal list of key indicators exists, and hundreds of indicators and data sources are of potential use. Given the extensive scope of the research project, data preparation activities form a distinctive phase of the project and need to be approached methodologically. The credibility of the research, as for any research, depends highly on the availability and quality of source data. The outcomes of this thesis will be deployed in subsequent stages of the research project and will contribute to the production of high quality research.

Firstly, data requirements and key data sources will be identified relying on the research background and economic theory. Relevant initial data shall be assembled into data repository that is designed and implemented to satisfy the criteria set by the nature of the research project. Data repository is designed so that it allows to produce structured, cohesive and systematically organized sets of data that are easily accessible for further econometric modelling and processing. In the pre-analysis phase a preliminary subset of indicators is proposed for further analysis through thorough feature selection process. As the last step, a set of tools for facilitating further pre-analysis (descriptive statistics) and data mining (conformity analysis) will be designed and implemented. These tools will provide overview of the data coverage and quality and facilitate the selection of final data sets for each econometric model. All these steps can be extremely time consuming and cumbersome in economic research that needs to use a lot of data from various sources and deals with various research questions. This thesis builds a coherent source of data for the international research team and enables them to filter out useful sets of data through automated processes and helps to avoid time-consuming data mining activities in econometric modelling phases.

The research project relies primarily on secondary data stemming from a wide array of public data sources (repositories of international and local institutions such as the World Bank, International Monetary Fund, World Trade Organisation, etc.). Research team aims to use various empirical research methods such as stochastic frontier analysis and dynamic panel data estimation techniques. The selection of specific research methods is significantly facilitated by this thesis.

The research project is led by Tallinn University of Technology (Estonia) and University of Lausanne (Switzerland) in co-operation with researchers from the National University of Laos (Laos), Ho Chi Minh City University of Law (Vietnam) and Royal University of Law and Economics (Cambodia) and is co-funded by the Horizon 2020 research grant No. 734712 „Institutions for Knowledge Intensive Development: Economic and Regulatory Aspects in South-East Asian Transition Economies“ (grant period 2017 - 2020).

The main beneficiaries of this thesis are the members of the research group, who can use the results of this thesis to continue with their research project. Additionally, the technical solution implemented in this thesis can be useful for anyone facing a complex research project that involves large sets of data that needs to be systematically structured and pre-analysed prior to the econometric modelling phase. This thesis can be especially valuable for those, who want to use World Bank databank as their primary data source.

1.1 Purpose

The central goal of the thesis is to build source data repository (database), which would serve as a single data source for the econometric modelling process that is part of the research project. The sub-objectives of the study are as follows:

1. To define data requirements and locate key data sources for the research project.
2. To design and implement a database solution to accommodate the source data.
3. To propose initial sub-set of relevant indicators for the research project.
4. To design and build tools which produce sets of data that is conformant with standard data analysis programs' requirements (standard panel data) and help to select final sets of data for (each) econometric modelling process.

This thesis needs to find answers to an array of sub-questions in order to achieve the objectives of the study. The study is faced with technical, methodological as well as design related problems:

1. What preliminary data is required to conduct the research project? What are key data sources?
2. What indicators (are used to) measure and characterize knowledge-based development and the role of institutional regime towards knowledge-based development?
3. What is the optimal technical solution (database platform) for data collection, storage and sharing?
4. How to design and implement the data repository (database)?
5. What are the most relevant indicators for the research project?
6. What kind of answers should the data pre-analysis tools be able to give? What measures are relevant?
7. How to enable data pre-analysis? What tools should be used?

Each of these sub-questions will be addressed in the suitable sections of the study.

1.2 Methodology

This thesis combines elements from economic research, system development, database design, data analysis and mining. It can also be viewed as a separate sub-process in the wider iterative econometric modelling process. Given the interdisciplinary nature of the study, no guideline methodology for conducting this thesis exists. The best practises from all listed domains have been implemented to the extent reasonable and applicable given the unique nature of the study. The data requirements have been identified based on the background of the research project, research hypotheses and thorough research into the knowledge economy assessment frameworks established by renowned international organisations and institutions such as the World Bank, Organisation for Economic Co-operation and Development (OECD) and Maastricht Economic and Social Research Institute (MERIT). The selection, design and implementation of the technical solution rests on the works of Connolly and Begg (2002) and Eesaar (2008).

The methods for the data pre-analysis have been selected considering the research project background, process and goals. Tools (incl. the algorithm) that enable the conformity analysis have been designed and implemented based on the works of Võhandu et al (2006) and Liiv et al (2007).

Author has formed a customized methodology (process map) for carrying out this thesis.

1.3 Overview of the study

The thesis contributes to a research project undertaken by consortium of five universities led by the research team from Department of Economics in Tallinn University of Technology. Econometric modelling process entails many phases and this thesis is a sub-process in wider iterative econometric process. The thesis is broken into three highly interrelated sections, each containing several sub-phases:

1. Definition of preliminary data requirements
    a. Investigate research background and objectives
    b. Identify generic data requirements and data sources
2. Design and implement source data repository
    a. Define and understand requirements to the database
    b. Investigate alternative solutions and select optimal solution

    c. Design of the database
    d. Implementation of the database
3. Data pre-analysis
    a. Conduct and describe feature selection
    b. Select and implement pre-analysis tools (descriptive statistics and conformity analysis)

It is important to note that this thesis is concentrated solely on specific objectives as laid out above and some aspects relevant to this thesis are excluded from the scope of this thesis:

1. Definition of research project's detailed hypotheses. The study summarizes the preliminary generic hypotheses of the research project based on the research proposal in order to shed light to the background of the research project and define data requirements, but the study is not concerned with the definition of the final research hypotheses (which is subject to further research and also to the outcome of this thesis).
2. Identification of specific sets of indicators for each econometric model. The thesis will propose only a preliminary pool of relevant indicators, which might be used (or not) in the research project (decided by the research team), and tools which will help to identify these final sets of data.
3. Interpretation of the results obtained through pre-analysis tools. This thesis will provide simple tools to run pre-analysis on the initial data, but it will not analyse the results. This thesis will however conduct initial feature selection and suggest the initial pool of indicators that should satisfy the data needs of the research project.
4. Selection and application of specific research methods (econometric modelling). The thesis does consider the generic requirements to data (format, structure etc.) relevant in the modelling process.
5. The motivation to conduct the research project is discussed briefly. This will be discussed in detail in separate upcoming publications (Hazak, forthcoming).

The study can be regarded as a client-contractor relationship, whereas the author of the study is the contractor and the research team is the client. As the roles of this relationship were highly interconnected, no real concerns usually associated with this kind of relationship were anticipated and therefore also not addressed in this thesis.

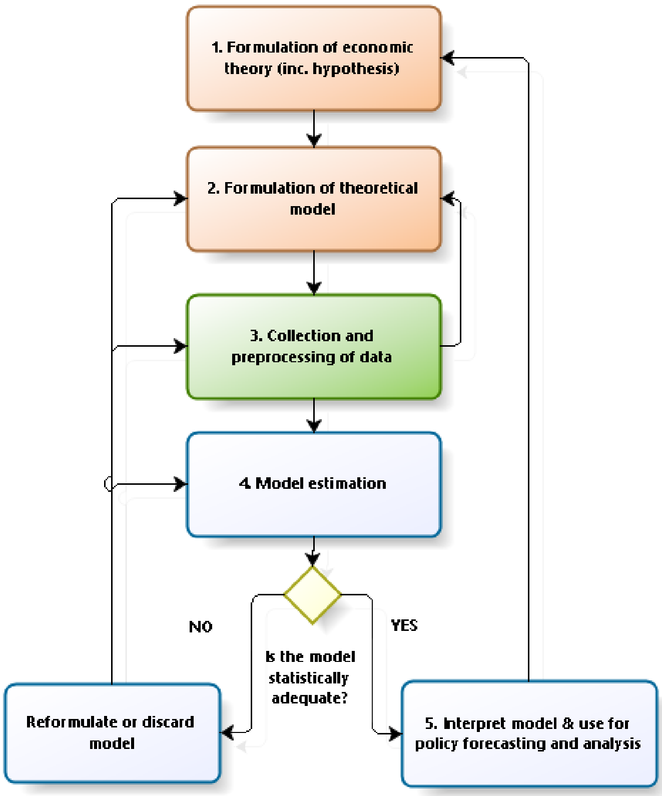
2 Methodology

At first, econometric modelling methodology will be briefly reviewed in order to locate and provide overview of the phases of the econometric modelling process that this thesis is related to. Next, an overview of a customized methodology, upon which the research questions will be based, will be introduced.

2.1 General methodology for econometric modelling

Standard econometric modelling process could be regarded in five inter-connected phases (See Figure 1). It starts with the formulation of the economic theory central to the study. Theoretical economic hypothesi(e)s are put forward based on the analysis of conceptual underpinnings and theoretical concepts relevant to the research. Next, specific problem(s) will be given mathematic form and translated into econometric model(s), followed by data collection and data processing activities, during which data is collected from various sources and prepared for modelling. After that the parameters of the econometric model(s) are estimated, giving empirical content to the defined functions. Model in general is evaluated from the standpoint of credibility and suitability in the context of the specific economic problem under study. If the model turns out to be inadequate the process returns to the beginning phases - either model needs to be reformulated and/or more data to be collected or a different estimation technique has to be applied. Once the model is satisfactory, the hypotheses are tested and the model is being interpreted and assessed in the context of the further practical usability. If the chosen model does not refute the hypothesis or theory under consideration, the model can be used for forecasting or prediction and also as a basis of political decisions.

Figure 1 Phases of econometric modelling based on (Paas, 1995) and (Brooks, 2008)


1. Formulation of economic theory (inc. hypothesis)
2. Formulation of theoretical model
3. Collection and preprocessing of data
4. Model estimation
NO
YES
Is the model statistically adequate?
Reformulate or discard model
5. Interpret model & use for policy forecasting and analysis

Econometric modelling usually involves several iterations, often caused by unavailability or low-quality data, forcing researchers to modify the scope of the work and/or to test special methods of estimation given unique nature of most economic data.

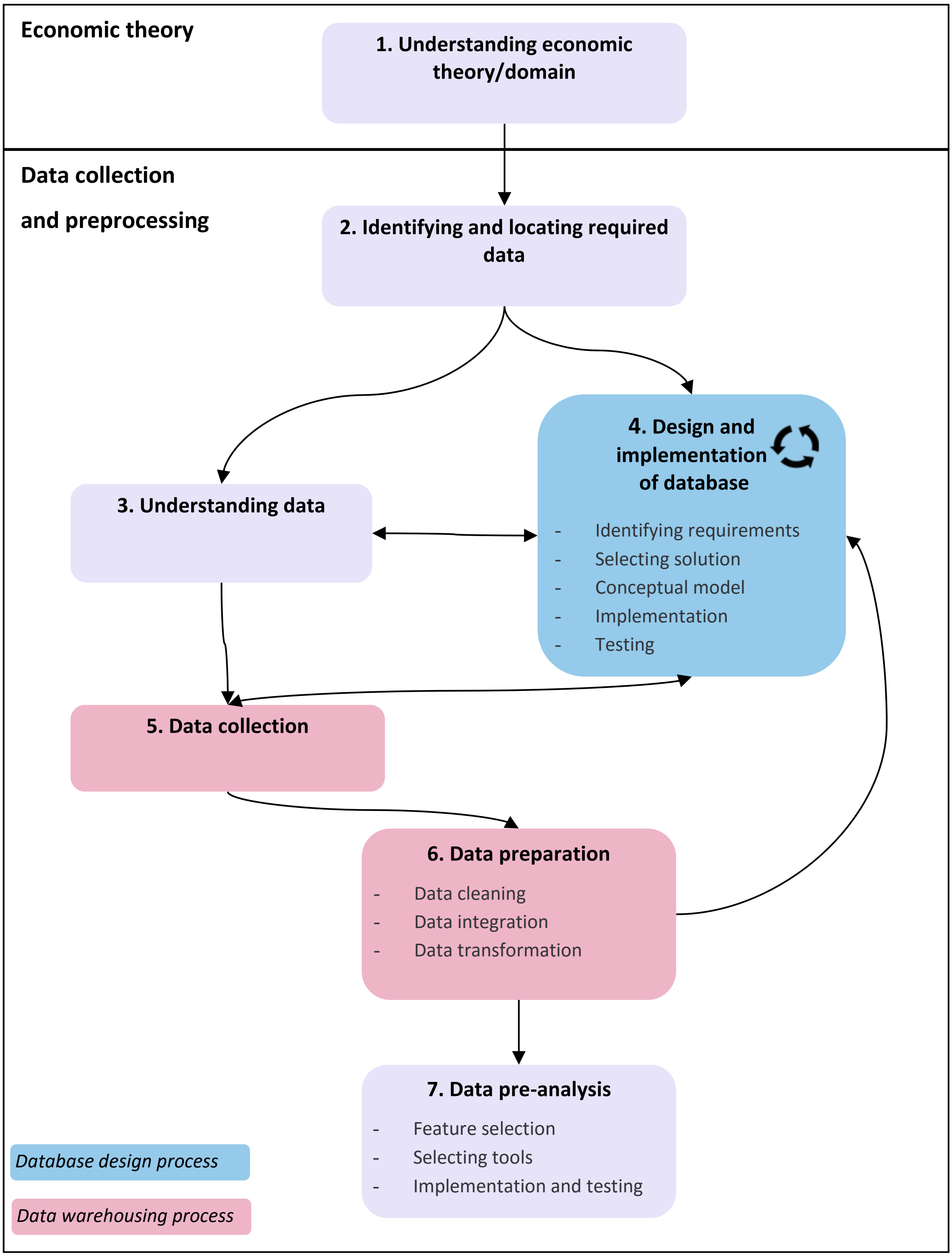
The phases this thesis is related to are coloured in red and green on the Figure 1. This thesis is primarily concerned with data collection and pre-processing phase (in red), which is of crucial importance for the entire process, but must also touch upon the preceding phases (in green), which essentially set the data requirements.

This thesis must seek to address to the extent possible and feasible the well-known shortcomings (problems related to small number of observations) of economic data, which is mostly secondary data.

2.2 Thesis methodology

As discussed above, this thesis can be viewed as a separate subprocess in the wider iterative econometric modelling process. Econometric modelling methodology provides a wide framework of how a standard economic research should be conducted but fails to provide guidance on how to specifically conduct this thesis. This thesis in its nature is an interdisciplinary project combining elements of economic research, data mining, data analysis, system development and database design. Hence no guideline methodology for conducting this kind of project really exists. Therefore, author has formed a customized methodology (process map) for carrying out this thesis. The graph below illustrates the process map of this thesis.

Figure 2 Thesis process map



First step is to understand the conceptual background, economic theory and research domain. Next step is to identify and locate required data. In this step, an overview of relevant data sources is obtained. Then, data is explored in order to establish how the data looks like, what data will be provided from the data sources, in which format it is shared, and how data can be accessed. This understanding is crucial to database design process. Database design and implementation will follow an iterative development process during which database is constructed incrementally. In the next phase, data is collected and loaded to the database followed by the data preparation phase. Collectively these phases can be juxtaposed to data warehousing process ETL (Extract, Transform, Load). During data preparation phase, several data management activities are carried out with the aim of producing cleaned and structured dataset suitable for further econometric modelling as well as data pre-analysis. While big part of the activities of the corresponding phases are carried out in the shown order, it is important to emphasise that steps 3-7 are highly interrelated. The last phase of this thesis aims to build a set of tools which enable to conduct simple initial comparative analysis of the data in the database. Through feature selection process an initial selection of relevant indicators is presented. The tools of the pre-analysis should enable comparative overview of presented indicators in terms of quality, availability and statistical strength across target countries and domains and help to design final subsets of indicators and countries. It is important to note that, in each step, dialogue with the research team is continuously maintained and feedback is considered during each phase of the study.

As this thesis deals with large amounts of data, it entails many elements of data mining. According to KDnuggets, a leading business analytics, big data and data mining webpage, one of the most popular data mining methodologies is CRISP-DM (Cross Industry Standard Process for Data Mining) (Piatetsky, n.d.). CRISP-DM methodology provides a model for the data mining project life cycle, drawing many parallels from software development life cycle. The main phases of the cycle are business understanding, data understanding, data preparation, modelling, evaluation and deployment. The standard accentuates the process' cyclicality and non-rigidness: the outcome of each process determines which phase or task needs to be performed next (Roman, 2016). Although this thesis does not complete full circle of data mining project (this thesis does not deal with modelling, evaluation and deployment), similar approach is adopted in this thesis in phases 3-7 (See Figure 2) that deal with data collection and preparation.

3 Data needs and data sources

One of the main tasks of this thesis is to design and build source data depository for the research project. The aim of this section is to define the data requirements of the data depository. In order to do that, first the economic theory along with the background of the research project will be investigated. The analysis of data requirements is further aided by the review of possible indicators, as proposed by other authors investigating institutional factors of knowledge based economic development. These activities will help to map the main data sources, understand the data quality (and issues with the quality) and identify sets of indicators subject to data collection activities.

3.1 Background of the data needs

There is an increasing acceptance of the idea that we are entering a new type of ‘knowledge economy’ (Smith, 2000). It is widely accepted that (application of) ‘knowledge’ has become one of the key drivers and the most critical resource of productivity and economic growth in present times. In broad terms, knowledge economy refers to the use of ‘knowledge’ to produce economic benefits (Günther, 2005). Most of the developed countries’ (countries belonging to OECD) economies today are knowledge-based, which means they are based on the production, distribution and use of knowledge and information, demonstrating high knowledge intensity (OECD, 1996). Less developed countries are on the path towards knowledge-based economy. According to Asian Development Bank (2007) wealth creation through application of human knowledge and creativity is steadily outpacing wealth creation through extraction and processing of natural resources. Thus, knowledge has increasingly become an important means for value creation.

The main aim of the research project is to investigate the economic development in South-East Asian countries towards the knowledge-based economy and society. The research project puts emphasis on understanding the role of different institutions[1] in knowledge intensive economic development of South-East Asian countries. Institutional efficiency and economic regime are considered as key determinants towards KBD (see below). The research project is concentrated on Cambodia, Laos and Vietnam – emerging market economies with shared history of wars, political instability and regime changes, which has left them far behind in economic development compared to other Southeast Asian countries. Regardless of their current remoteness from the global knowledge economies, these countries are well positioned to exploit the momentum provided by the economic transition to set themselves on the fast track towards knowledge-based economy (Asian Development Bank, 2014).

Hazak (forthcoming) asserts that the prioritisation and deployment of knowledge within an economy remains a key success factor for long term economic development. Econometric tests run by the World Bank (2008) demonstrate a statistically significant causal relationship running from the level of knowledge accumulation, as measured by Knowledge Economy Index (KEI), to future economic growth. Hence, productivity and growth are becoming increasingly dependent on knowledge and knowledge based industries. The research project aims to explore the micro and macro level mechanisms that encourage knowledge creation and absorption in parallel with the investigation into the mechanisms and incentives that aid to transfer knowledge into lasting value within Southeast Asian context (Hazak, forthcoming).

Various international organisations and institutions such as The World Bank, OECD, and Asian Development Bank (ADB) along with numerous scientists, researchers and policy makers seem to agree on the main pillars of KBE. The central concept of KBE is that favourable economic and institutional environment along with the sustained investments in education, innovation systems, ICTs and infrastructure will pave the path to increased

[1] In wide context, institution can be defined as “established law or practise” (Oxford Dictionaries, 2017). North (North, 1990) defines institutions as “the rules of the game in a society”, which are human devised constrains that shape the human interaction. Institutions, in the context of this thesis and the research project, refer to the various formal and informal mechanisms and structures of social order such as government, economic and legal systems, educational institutions, research community, family, religion etc., that govern the behaviour of individuals. The research project will primarily be interested in formal institutions and will use number of indicators that characterize these institutions.

creation and application of knowledge in economic production which in turn leads to economic growth (Chen & Dahlman (2005), Asian Development Bank (2007), Powell & Snellmann (2004), OECD (1996 and 2001)).

Among all these elements, government and broader institutional environment plays pivotal role as it holds the capacity to induce favourable regulatory climate for innovation, business and entrepreneurship, to create adaptive and inclusive labour markets and to promote the investment into R&D and ICT infrastructure. The institutional regime has an equally vital role to play in coordinating and linking the various efforts in the economy as all the pillars of KBE are highly interrelated.

There is ample research and evidence from developed countries which suggests that well-functioning institutions are crucial to (knowledge-based) economic development (such as Timmer (2006), North (1990), Acemolu, et al (2005)). Institutional accountability, enforcement of contracts, rule of law, freedom of speech and property rights are preconditions among many others that must be established by the institutional regime in order to attract investments, reduce transaction costs and set ground for economic growth (Timmer, 2006). Corruption, fraud, red tape, regime uncertainty and lobbying among many others on the other hand are found (Mo, 2001; Ehrlich and Lui, 1999) to be the key institutional inefficiencies that halt economic growth. Hazak (forthcoming) argues that these challenges are especially important in the transition economies such as Cambodia, Laos and Vietnam which sooner or later will need to revise their institutional and regulatory environment.

Institutions play also major role in developing national innovation systems. National innovation systems (networks of universities, private and public research institutions and think tanks), determine the ways in which innovation and knowledge is acquired, created, disseminated and applied (Chen & Dahlman, 2005). Favourable regulatory climate encourages interactions among the different innovation system players (universities, private and public research institutions, think tanks) (Asian Development Bank, 2007). Furthermore, institutions have the capability to incentivize the investment into knowledge, high-tech and human capital intense industries and to reduce the risks and uncertainties associated with these new fields of economic growth. Several studies (Lederman and Maloney (2003), Guellec and Van Pottelsberghe de la Potterie (2001), Griffith et al (2004)) have convincingly demonstrated the positive effect of innovation (induced by investment into R&D and R&D intense industries) on economic and productivity growth.

The success of cross regional knowledge and technology transmission as well as diffusion is highly dependent on county's absorptive capabilities such as the level of human capital and IT infrastructure (Hazak, forthcoming). Institutional regime needs to improve equal access to and quality of education, which are critical in building skilled and technology savvy workforce that knowledge-based development relies on. The positive correlation between the level of education of a population and economic growth is well documented by Barro (1991) and Cohen & Soto (2001). Additionally, governments must also develop and grant equal access to ICTs, which will provide access to global knowledgebase and networks. Finding balance in the liberalization and deregulation whilst promoting the use and development of ICTs is one of the current challenges of developing Asian economies (Asian Development Bank, 2007).

In general, there is very limited research on the role of institutions and regulations in Cambodia, Vietnam and Laos on the path towards higher knowledge intensity. Most of the studies on research, knowledge and high-technology based growth have focused on developed countries. According to the latest World Bank country report (World Bank Group, 2017), Laos has in recent history made significant advances in the development by improving access to education, health, and infrastructure, decreasing poverty and increasing incomes. Worryingly, most of the GDP growth is still driven by natural resources and little value added is generated by modern industries such as financial sector and ICTs. World Bank concludes that strengthening institutions and enhancing government is key to further progress. Similarly to Laos, Cambodia has also demonstrated steady economic growth since recession, yet long term growth is threatened by low competitiveness embodied in form of weak institutions, poor infrastructure, low quality of education and lack of innovation stimuli (World Bank Group, 2017). Vietnam, named as "one of the world's great development success", needs to build a more competitive private sector, promote innovation, and tap into trade opportunities to carry out broad structural reforms (World Bank, 2016).

In these transition economies, institutional and regulatory inefficiencies seem to be detrimental obstacles on the route towards knowledge-based development (Hazak, forthcoming). The research project is very valuable since it complements the existing limited international as well as regional literature and research on transition of Southeast Asian economies towards KBE.

3.2 Hypotheses

The project seeks to understand the role of different institutions in transition economies, with focus on Cambodia, Laos, Vietnam, in the process of moving towards KBE. The project aims also to investigate the causes and differences among these three countries in the KBD.

The main research goal of the project is to provide a better understanding of the role of institutional mechanisms towards KBE as well as regulatory incentives and measures (i.e. those addressing market frictions and inefficiencies such as transactions costs, taxes, agency and information problems etc) that could be employed to accelerate the transition (Hazak, forthcoming).

These research goals will be reached through iterative econometric modelling process. Based on the previous research the research team along with the author have defined preliminary set of general hypotheses (subject to possible revisions and specifications contingent on availability of data), that are subject to testing with data collected during this thesis. Some of the preliminary set of core hypotheses are as follows (Hazak, forthcoming):

(H1) Certain knowledge capturing capabilities are key drivers towards a knowledge economy.

(H2) Certain institutional factors (e.g. level of education, competition, corruption) and financial incentives (e.g. access to capital markets, risk profile of knowledge intensive investments) influence the transmission of knowledge into economic value across countries and explain the cross-country differences.

(H3) Knowledge-based (capital) investments depend on the individual, company or country level asset/income structure.

(H4) Differences in the regulatory framework are among the key drivers of the differences in the knowledge intensity of countries and industries.

(H5) Regulatory measures help to reduce obstacles (such as market frictions, insufficient investment protection and credit constraints) for knowledge related investments.

3.3 Indicators for knowledge-based economy

Given the vast scope of knowledge economy, the topic of knowledge economy indicators is equally complex. Although major efforts have been made in the field of innovation indicator development in order to develop better quantitative indicators for innovation (e.g. knowledge), knowledge measurement and knowledge economy remains a key challenge (Smith (2000), OECD (1996)).

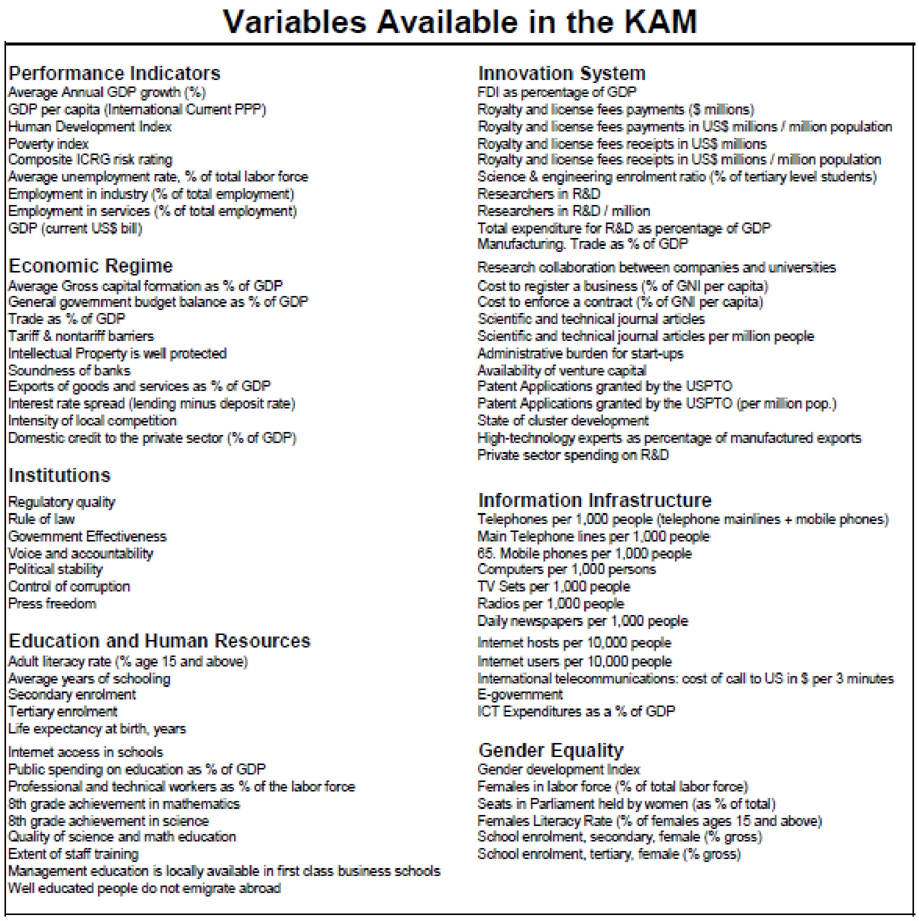
The main problem is that knowledge itself is particularly hard to price and to quantify; unknown proportion of knowledge is implicit, uncodified and stored only in the minds of individuals (OECD, 1996). Therefore, knowledge and the knowledge-based economy can be measured only via indirect indicators. Although the transition of global economy to a KBE, led by innovation, is widely recognized, given the complexity surrounding the measurement of knowledge and knowledge economies, no universal list of indicators for mapping and measuring the KBEs exist. Hence, to determine the initial pool of variables important in the context of the research project various knowledge economy assessment frameworks established by international organisations and institutions will be examined.

3.3.1 World Bank Knowledge Assessment Methodology

Knowledge Assessment Methodology (KAM) is a widely used framework developed by the World Bank as part of the Knowledge for Development Program. The program is designed to provide a basic assessment of countries' readiness for the knowledge economy and to identify sectors or specific areas that are hindering the development (Chen & Dahlman, 2005). KAM framework also allows countries to assess how they compare with others in their ability to compete in the global knowledge economy. According to the framework, the four pivotal pillars required for successful transition to the knowledge economy are (Chen & Dahlman, 2005): 1. sustained investments in education, 2. development of innovation capability, 3. modernization of the information infrastructure and 4. creation of a conducive economic environment and institutional regime.

The most recent KAM (2008) builds on 83 structural and qualitative variables (see table below) that serve as proxies for the four knowledge economy pillars. The framework allows for four different modes (global scale, regional scale, basis of human development, basis of income levels) of comparative assessment of the relative performance of countries and regions on the knowledge economy (The World Bank, 2008). Variables are normalized from 0 to 10 (strongest) and ranked on ordinal scale.

Figure 3 Variables of KAM



Variables Available in the KAM

Performance Indicators
Average Annual GDP growth (%)
GDP per capita (International Current PPP)
Human Development Index
Poverty index
Composite ICRG risk rating
Average unemployment rate, % of total labor force
Employment in industry (% of total employment)
Employment in services (% of total employment)
GDP (current US$ bill)

Economic Regime
Average Gross capital formation as % of GDP
General government budget balance as % of GDP
Trade as % of GDP
Tariff & nontariff barriers
Intellectual Property is well protected
Soundness of banks
Exports of goods and services as % of GDP
Interest rate spread (lending minus deposit rate)
Intensity of local competition
Domestic credit to the private sector (% of GDP)

Institutions
Regulatory quality
Rule of law
Government Effectiveness
Voice and accountability
Political stability
Control of corruption
Press freedom

Education and Human Resources
Adult literacy rate (% age 15 and above)
Average years of schooling
Secondary enrolment
Tertiary enrolment
Life expectancy at birth, years
Internet access in schools
Public spending on education as % of GDP
Professional and technical workers as % of the labor force
8th grade achievement in mathematics
8th grade achievement in science
Quality of science and math education
Extent of staff training
Management education is locally available in first class business schools
Well educated people do not emigrate abroad

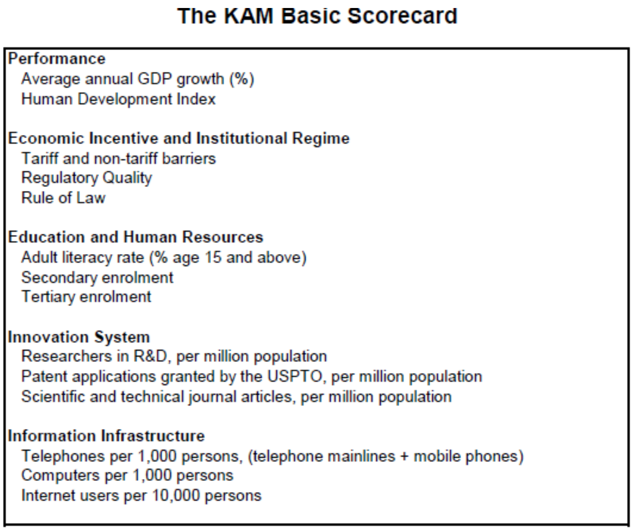
Innovation System
FDI as percentage of GDP
Royalty and license fees payments ($ millions)
Royalty and license fees payments in US$ millions / million population
Royalty and license fees receipts in US$ millions
Royalty and license fees receipts in US$ millions / million population
Science & engineering enrolment ratio (% of tertiary level students)
Researchers in R&D
Researchers in R&D / million
Total expenditure for R&D as percentage of GDP
Manufacturing. Trade as % of GDP
Research collaboration between companies and universities
Cost to register a business (% of GNI per capita)
Cost to enforce a contract (% of GNI per capita)
Scientific and technical journal articles
Scientific and technical journal articles per million people
Administrative burden for start-ups
Availability of venture capital
Patent Applications granted by the USPTO
Patent Applications granted by the USPTO (per million pop.)
State of cluster development
High-technology experts as percentage of manufactured exports
Private sector spending on R&D

Information Infrastructure
Telephones per 1,000 people (telephone mainlines + mobile phones)
Main Telephone lines per 1,000 people
65. Mobile phones per 1,000 people
Computers per 1,000 persons
TV Sets per 1,000 people
Radios per 1,000 people
Daily newspapers per 1,000 people
Internet hosts per 10,000 people
Internet users per 10,000 people
International telecommunications: cost of call to US in $ per 3 minutes
E-government
ICT Expenditures as a % of GDP

Gender Equality
Gender development Index
Females in labor force (% of total labor force)
Seats in Parliament held by women (as % of total)
Females Literacy Rate (% of females ages 15 and above)
School enrolment, secondary, female (% gross)
School enrolment, tertiary, female (% gross)

According to World Bank (The World Bank, 2008), the most used modes of KAM are their Basic Knowledge Economy Scorecards and Knowledge Economy Index (KEI). Both rely on 14 standard variables (see table below), of which two are performance variables and 12 knowledge variables representing the four pillars of knowledge economy. These 14 variables may be viewed as core indicators of knowledge economy that are generally available for large time series and remain regularly updated for vast majority of countries (Chen & Dahlman, 2005). Methodologically, KEI is constructed as simple average of the normalized performance scores of a country or a region on the key variables in four knowledge economy pillars, summarizing the performance over the four KE pillars for a country or a region (The World Bank, 2008). Basic scorecard can be thus seen as a disaggregated representation of KEI.

Figure 4 World Bank KAM Basic Scorecard



The KAM Basic Scorecard

Performance
Average annual GDP growth (%)
Human Development Index

Economic Incentive and Institutional Regime
Tariff and non-tariff barriers
Regulatory Quality
Rule of Law

Education and Human Resources
Adult literacy rate (% age 15 and above)
Secondary enrolment
Tertiary enrolment

Innovation System
Researchers in R&D, per million population
Patent applications granted by the USPTO, per million population
Scientific and technical journal articles, per million population

Information Infrastructure
Telephones per 1,000 persons, (telephone mainlines + mobile phones)
Computers per 1,000 persons
Internet users per 10,000 persons

World Bank claims (The World Bank, 2008) that the data on which the KAM is based are all published by reputable institutions that are at the forefront of gathering and producing reliable and internationally consistent country statistics.

3.3.2 OECD framework

OECD is one of the main investigators of KBD in the developed countries and has had a significant role in the development of knowledge economy indicators. One of its first attempts to compile a comprehensive set of statistical indicators relevant for knowledge economy dates to 1996, when it published a landmark report “The Knowledge-based Economy” (OECD, 1996) by being one of the first international institutions to recognize the growing importance of ICTs and its impact on economic development. Few years later, it initiated the “Growth Project” with the aim of exploring the underlying causes of differences in growth performance in the OECD area over the preceding decade (OECD, 2001). The final report concluded, that while ICT has indeed led to more rapid growth in some countries, “growth is not the result of a single policy or institutional arrangement, but a comprehensive and co-ordinated set of actions to create right conditions for future change and innovation". It encouraged the countries to adopt comprehensive growth strategy emphasizing:

- Macroeconomic stability, openness and effectively functioning markets and institutions;
- Diffusion of ICTs;
- Fostering innovation by prioritizing fundamental research, improving funding for public R&D and promoting flow of knowledge between science and industry;
- Investing in human capital; and
- Stimulating firm creation.

OECD has also expressed concerns over the quality and validity of knowledge economy indicators that are widely used in the context of knowledge economy. This critique will be discussed later.

3.3.3 Knowledge Economy Indicator project by MERIT

Maastricht Economic and Social Research Institute on Innovation and Technology (MERIT) is a research and training institute of United Nations University. Process of technological change and innovation in global perspective is at the focus of MERIT's research. In 2008 it carried out the Knowledge Economy Indicator project, that purpose was to identify key indicators for knowledge economies and methodologies for constructing composite indicators to measure and compare national KBE performance (Arundel, Hansen, & Minna, 2008).

Hundreds of indicators were evaluated for their usefulness in evaluating and tracking the development of KBE in Europe and among many other countries outside Europe. Report summarizes 64 key indicators, which were classified as drivers, characteristics and key outputs of KBE. The authors of the research emphasise that, although hundreds of indicators are of potential use then many suffer problems of availability and consistency.

Proposed indicators, classified as drivers and characteristics of KBE, are grouped under following four central themes (Arundel, Hansen, & Minna, 2008):

- Production and diffusion of ICTs;
- Human resources, skills and creativity, as means of advancing the creative and absorptive capacity of a work force;

- Knowledge production and diffusion. This subgroup includes indicators mostly on R&D activities;
- Innovation, entrepreneurship and creative destruction. These elements demonstrate the change brought about by ICTs and globalising knowledge economies (e.g. demand for innovative products).

Group B indicators include next to economic output indicators also measures on social performance and quality of life:

- Economic output;
- Social performance. Indicators characterizing the environment and sustainable growth, economic welfare and quality of life.

Arundel, Hansen, & Minna (2008) emphasizes the need to move beyond the traditional indicators of KBE and therefore add number of KBE concept expanding measures (under Group C) in areas of economics and work life, trade, knowledge production and diffusion, economic structure and human resources.

3.3.4 APEC framework

The last framework examined is the one offered by the Asia-Pacific Economic Cooperation (APEC) in early 2000's. The project's aim was to provide analytical basis that would be useful in promoting the effective use of knowledge, and creation and dissemination of knowledge among APEC economies (Asia-Pacific Economic Cooperation, 2000). APEC studied a representative range of APEC countries across select set of characteristics and indicators relating to the development towards KBE and identified characteristics that are preconditions of KBE.

The quantitative indicators used in APEC study attempt to capture the general stage of development of these economies relative to a fully developed knowledge-based economy and the economies' current potential to become KBEs. Indicators cover following groups of characteristics: (1) innovation system, (2) human resource development, (3) ICT infrastructure, and (4) business environment.

As an interesting point, APEC warns that there are many indicators measuring some characteristics of KBE, but few indicators which actually measure the extent to which a country is already operating as a KBE (Asia-Pacific Economic Cooperation, 2000).

APEC suggests looking at a proportion of current economic activity that is in some sense “knowledge intensive”. Knowledge intensity could be measured either via by money or by the number of people involved in knowledge intense industries.

3.3.5 Critique of KBE indicators

Measurement of economy has always been challenging, but even more so for the KBE. Current traditional economy indicators (which focus on aggregate values of goods and services and are designed for traditional economy) may fail to capture the fundamental aspects of economic performance to the extent to which KBE differs from traditional economic theory (OECD (1996), Smith (2000)).

The four main reasons why knowledge indicators cannot approximate the systematic comprehensiveness of traditional economic indicators are as follows (OECD, 1996):

- Even though knowledge will generally increase economic output, the effect on economic output in qualitative and quantitative terms is unknown in advance;
- There are no intellectual capital accounts (e.g. knowledge) analogous to fixed capital accounts in the national account systems, which makes it hard to map knowledge inputs;
- The absence of systematic price information does not allow to aggregate individual knowledge transactions into broader aggregates;
- New knowledge creation is not necessarily net addition to knowledge stock, as it may render some old knowledge obsolete.

In order to capture KBE, one needs to measure knowledge inputs, stocks and flows, outputs, networks, knowledge and learning. The main problems surrounding the application of knowledge indicators and measurement of KBE:

- Inability to correctly identify indicators as inputs or outputs (OECD, 1996);
- Much of the KBE measurement is input focused (OECD, 1996);
- The expenditure on R&D is over emphasised as an input to knowledge production; only small amount of R&D counts for total knowledge creation and it should not be treated as a single input to knowledge production. The further implication of

this is flawed classification of companies based on R&D expenditure into clusters of low-, medium- or high-tech companies (one of the main metrics characterising the knowledge intensity of companies and countries, published by OECD) (Smith, 2000).

- On country level, R&D indicators tend to account only for spending incurred by public sector or large manufacturing companies, dismissing the R&D incurred by service sector and small firms (OECD, 1996);
- Patents are regarded as one of the best ways to measure knowledge production but not all patents are equally significant, nor all new applications of knowledge are patented (OECD, 1996). Moreover, the number of patents as such tells very little about the economic impact of the invention (Smith, 2000);
- Knowledge flows and stock are particularly hard to measure due to minimal transaction information (Smith, 2000);
- In context of measuring the absorptive potential of human capital, often PISA scores in maths are used to characterize the aptitude of human capital. However, based on the general theory of KBE, skills such as reading, creativity and communication skills are equally vital for knowledge workers (Arundel, Hansen, & Minna, 2008). Thus, indicators based solely on mathematical skills may fail to capture the level of human capital in a country.

3.4 Synthesis of KBE framework analysis

The review of the frameworks revealed that the frameworks tend to evolve around very similar concepts: quality of human capital/education, innovation system, (ICT) infrastructure, business environment and general economic performance and institutional regime (see Table 1 below).

Table 1 Comparison of KBE measurement frameworks

Framework	**World Bank KAM**	**OECD**	**APEC**	**MERIT**	
Aim	Tool to assess country's development towards KBE 83 indicators	Measurement of KBE	Assess level of development compared to fully developed KBE 25 indicators	Measurement of KBE, methodology to construct composite index 64 indicators	
Indicator clusters	Performance (incl. human development index)	Macroeconomic stability, effective markets and institutions		Economic output, social performance	Globalization indicators
	Education and human resources	Investment in human capital	Human resource development	Human resources, skills and creativity	
	Innovation system	Fostering innovation system	Innovation system	Knowledge production and diffusion	
	Information infrastructure	Diffusion of ICTs	ICT infrastructure	Production and diffusion of ICTs	
	Economic regime and institutions	Stimulating firm creation	Business environment	Innovation, entrepreneurship and creative destruction	

The colouring indicates the relative overlapping of themes across the frameworks.

Although frameworks allocate different weights to abovementioned clusters and the categorization of indicators might differ slightly, all clusters are represented in all four frameworks to smaller or greater extent. Table 2 highlights some key indicators for each pillar of KBE.

Table 2 Sample indicators of KBE

Quality of human resources/education	**Innovation system**	**Infrastructure/Diffusion of ICTS**
Adult literacy rate %	Researchers in R&D	Internet users
Secondary enrolment rate %	Patent applications granted	Telephone users
Tertiary enrolment rate %	Patent applications submitted	Computer users
Human development index	Scientific and technical journal articles	E-government
Public spending on education as % of GDP	R&D expenditure as % of GDP	
Institutional efficiency	**Business environment**	**Economic performance**
Rule of law index	FDI as % of GDP	GDP growth %
Regulatory quality	High-tech exports	GDP per capita
Government transparency rating		GDP
Government effectiveness rating		
Press freedom		
Corruption index		

KAM, OECD and APEC frameworks are analogous, only APEC framework not taking in any indicators that measure general economic performance. MERIT has the most focused and complex view to the measurement of KBE. It puts a lot of emphasis on measuring knowledge production in terms of inputs (as different modes of R&D expenditure of GDP) and outputs (different kinds of patents, research co-operations). It also includes indicators measuring the demand for innovative products and market innovation outputs. Surprisingly, the framework does not include any indicators relating to the measurement of institutional regime and effectiveness. Although MERIT's indicators are arguably most specific and effective in measuring the extent of KBE, most of the data sources for these indicators are only available for countries belonging to OECD. Hence, little of the indicators can be used for this thesis.

The analysis of frameworks suggests that literally hundreds of indicators are of potential use when analysing knowledge-based economies and development towards it. The selection of indicators is much more abundant for developed countries (countries belonging to EU, OECD), whereas data quality and availability issues concerning developing and less developed countries might significantly limit the number of indicators suitable for the analysis of knowledge intensity.

The research on KBE indicators enabled to:

- Identify the pool of adequate and available indicators used by established international institutions in the research on KBE and KBD. World Bank KAM framework, consisting 83 indicators, including extensive set of institutional indicators, serves as the best starting point for data collection activities;
- Identify potentially useful data sources for this thesis. Global institutions should be preferred to ensure data comparability and quality. World Bank has the most comprehensive datasets in terms of country and topic coverage;
- Map the pool of useful indicators to various dimensions of KBE;
- Understand which KBE indicators characterise inputs, outputs and knowledge flows of KBE;
- Take note of the pitfalls and problems concerning the indicators necessary for KBE and KBD analysis;
- Structure the process of indicator/data collection, organisation and recording.

3.5 Identified data requirements

Data requirements have been identified based on several considerations. Although the conceptual underpinnings, goals of the research and research hypotheses define the main data requirements, the analysis of various KBE frameworks has proven to be equally useful and informative in setting the data requirements.

Author in collaboration with the research team considering the theoretical framework of the research project and analysis conducted on the KBE frameworks has identified following data requirements:

1. The main objects about which data is collected is ‘country’. Data regarding all major world countries (as per World Bank) will be collected in order to enable comparative analysis of knowledge intensity and KBD across Southeast Asian countries as well as on select sample economies outside Asia. In order to enable more meaningful analysis, industry and company level data also is highly desirable. However, based on the initial review of potential data sources, the availability of such data in comparable format across countries is very poor. Thus, most likely ‘country’ will be the main level of data collection.

2. The indicators (variables) regarding following categories are sought after:

   - Structural – Indicators providing descriptive information regarding countries (such as land area, arable land area, religion, etc.). These indicators can be used as control variables in econometric modelling.

   - Demographics – Indicators describing the country’s demographics (population density, rural/urban population, age profile etc.).

   - Human development indicators will be split into two groups:

     - Public health – Indicators measuring the quality of life and the well-being of the citizens (birth rate, life expectancy, health expenditure data).

     - Education – Indicators defining the quality and capabilities of human capital (school enrolment rates, literacy rate, PISA scores, government expenditure on education).

   - Economic performance – Indicators describing country’s level of development, economic output and the structure of the economy (GDP per capita, interest rate, real GDP growth, services/agriculture/industry value-added, index of globalization).

- Innovation system (knowledge intensity) – Various measures describing the state of country's innovation system, that is its ability to initiate, import, modify, and diffuse new technologies and practices (high-technology exports, R&D expenditure as % of GDP, trademark/patent applications).
- ICT infrastructure – Indicators demonstrating the ability of the citizens and businesses to diffuse knowledge and access global knowledge-base.
- Business environment - Various metrics measuring the ease of doing business in a country (capital requirements, legal procedures to start a business, tax system etc.).
- Institutional regime and efficiency – Various indicators describing the economic and legal policies of government, country's attractiveness for international investors, and its supportiveness for innovation and firm creation.

No input-output classification shall be made since this classification can be very subjective as demonstrated by the analysis of frameworks, that classified indicators very differently. The backbone of the data repository will be built on World Bank database that has the most comprehensive database among all international institutions.

To address the shortcomings of economic data, data should be collected over long periods and the database should include as much additional information regarding the collected data as possible (method of collection, sources, definitions etc).

4 Designing and implementing database

The next phase of the thesis deals with the design and implementation of the research database. The research database will hold data collected from various sources in a semi-structured and easily manipulatable format where it can be imported to data analysis programs such as Eviews, R or STATA.

But why this research project needs a database? Normally simple data analysis and manipulation tools such as MS Excel are sufficient for simple data pre-processing and structuring tasks needed to be undertaken prior to the econometric modelling phase in data analysis software (in 64-bit Windows environment, MS Excel file does not have hard size limits; however, one sheet is limited by ca 1 million rows and ca 16 thousand columns). Data analysis and statistics software tools also provide functionality to clean, reorganize, manipulate and overwrite data.

This research project needs database mostly for the following reasons:

- Preliminary dataset is expected to hold large amounts of data; over 200 variables/indicators across ca 215 countries over long period of time (depends on the availability of data). Such amount of data will have very low comprehensibility and visibility when processed directly in data analysis software;

- Data must be structured and systematically organized in order to enable the evaluation of data quality, availability and general suitability for the final econometric model(s) already during the pre-analysis phase;

- Data from multiple data sources of different formats must be combined into a unanimous format in order to enable data analysis;

- It must be possible to easily modify the preliminary dataset (opt in and out variables and countries) – database will be a “tool” that will help to model the final data sets used in econometric modelling phase;

- Pre-processed and structured preliminary dataset will ensure equal quality and format of the input economic data across various research teams and (their) economic models;

- To the extent reasonably feasible, data should be updatable as the research project is expected to last for minimum of 4 years. New data points are likely to become available during this period.

Drawing from above, some form of “database” is necessary in order to facilitate the econometric modelling process and foremost save time on data preparation activities that tend to consume a lot of time during econometric modelling process. Given the complexity of the research domain, feature selection is likely to occur in several iterations. Final data sets will be subject to many factors, such as availability and quality. The database will act like a “tool” that will help to visually gauge and systematically analyse the vast amount of information potentially useful for the research project and model the final data set(s). Database would contain all the necessary data in an organized, structured, modifiable and to the extent possible updateable format and would serve as a single data source for the modelling process.

4.1 Database type

In the context of this thesis it is important to clarify some of the terminology relating to databases. The term “database” has numerous meanings and definitions. In broad context, it can be viewed as an umbrella term for any sort of collection of data. However, in the field of information technology term “database” normally refers to a database administered with database management system, which is a collection of programs that enables users to create and maintain a database (Elmasri and Navathe, 2010).

By and large, databases are classified by data model (relational, hierarchical, network, object-orientated, XML etc.), by database distribution model (centralized, distributed) and by the usage purpose (on-line transaction processing (OLTP) vs on-line analytical processing (OLAP)) (Elmasri and Navathe, 2010). OLTP database systems, where data is detailed and current, are designed to support large number of simultaneous transactions, with the aim of making transactional systems run efficiently. Main functions of these databases are retrieval, update and deletion of single fact. In contrast, OLAP systems, characterized by low volume of transactions, are designed for analytic purposes. These systems support strategic and tactical decisions and deal with historic data. Data is stored normally in multi-dimensional star schemas. Main functions in these databases are extraction of large amounts of data and processing of complex queries. OLAP systems can also be called data warehouses or data marts.

In recent years, along with the growth of data-rich environments a term “data lake” has emerged, referring to “a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data" (Dull, 2017). Data lakes are highly agile, mostly unstructured storage repositories; data structure is not defined until the data is actually needed, enabling its users to easily reconfigure their models and queries (Dull, 2017). Data warehouse in contrast is a highly-structured system. Data warehouses are optimized for business professionals seeking answers for simple business questions, while data lakes are most useful for data scientists looking to solve more complex problems (Dull, 2017).

Database built during this thesis cannot be categorized to any specific database type described above and is not designed by following any strict design methodologies associated with the above-described database systems. This database is not designed to support any business environment and hence does not represent a highly structured environment. Database designed and built during this thesis can be viewed as a custom-built tool designed to solve specific problem in the context of the research project. At best, it can be viewed as an OLAP system resembling most to a data lake format, but it does not take the form of any specific database type. As such a collective term "database" has been used throughout this thesis.

4.2 Requirements for the database

There are several options to implement the database. The basis for the choice of the technical solution of the database is dictated by the (functional and non-functional) requirements that the solution must satisfy. Requirements are derived foremost by the needs of the research team and the parameters of the source dataset. The aim of this thesis is to design and implement a most optimal solution to the problem.

Author in collaboration with the research team has identified that the solution must satisfy following basic requirements:

1. Be either free open-source software or belong to the MS Office family;
2. Enable direct data import via World Bank Application Programming Interface (API);
3. Enable direct data export to data analysis software such as STATA/Eviews/R/MS Excel;
4. Hold minimum of ca 200x217x60 rows of data, min 1 GB of data.

5. Must be programmable (requesting and transforming data);
6. Enable to easily insert, delete and modify data;
7. Must be easy and intuitive to use, ample online documentation and support should be available;
8. Must be able to perform complex calculations on big sets of data in reasonable timeframes.

Database will be stored on researchers' private computers, which eliminates the need for a server-based solution.

4.3 Selection of the technical solution

Based on the requirements, feasible options are:

1. Database built into spreadsheet applications (MS Excel);
2. Flat file database, operated programmatically;
3. Relational database management systems (RDBMS).

MS Excel provides technically all functionality, but in not the most optimal way. It is rather difficult and inconvenient to combine data from several tables and create multi-layered queries with MS Excel. It might also have some performance issues due to the amount of data that will be stored. Although Excel file does not have hard size limits (in 64-bit Windows environment), one sheet is limited by ca 1 million rows and ca 16 thousand columns. MS Excel also allows users to design tasks with VBA. MS Excel most certainly is the simplest and easiest solution (for users), but likely performance shortcomings will eliminate this option.

It is also possible to create a database in a programmatic way. The output of such solution would be a flat file that can be read by data analysis software. Although all the functionality is met by such option, it requires a lot of programming capacity and would take long time to build. This solution would also require building a simple interface to communicate with the user. Although feasible, this is most certainly not the most suitable solution.

RDBMSs are software applications designed to manage databases. RDBMS are based on relational data model, where data is logically structured within relations (tables). RDBMS provide four main functionalities; data definition, update, retrieval and administration. There are many types of RDBMS ranging from simple solutions (such as MS Access and Filemaker) that run from personal computers to large systems (such as Progress, MySQL, PostgreSQL) that run on mainframes.

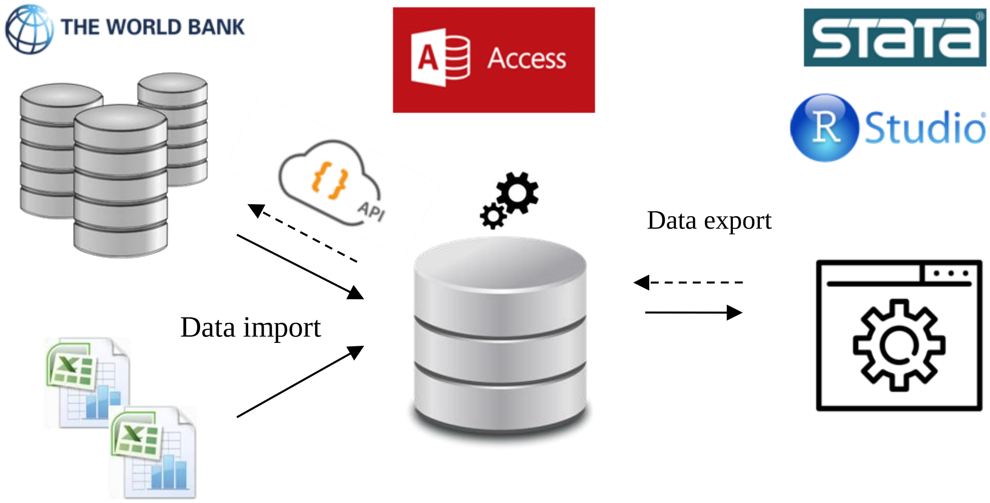
In the context of this project, it would be sensible to explore one of the most widespread simple desktop RDBMS - MS Access. MS Access offers all the core functionality necessary to manage databases through a simple graphical user interface - its major plusses, especially for those used to work in MS Excel. MS Access is useful tool for storing, sorting and retrieving data for variety of applications. It is built on relational Microsoft Jet Database engine and can hold up to 2 GB of data, which usually satisfies the capacity requirement. Like other MS Office tools MS Access provides tools to develop customized database applications using Microsoft Visual Basic for Application (VBA) language. As it is part of the MS Office package, it is already available on researchers' personal computers or on their university computers.

Given the above MS Access is the most optimal choice as it offers all the functionality needed and best usability.

4.4 Description of the technical solution and data flows

The figure below illustrates the architecture of the selected technical solution. The central component of the solution is a local MS Access database that will hold data from various data sources. The main data source is World Bank databank, from where majority of the data is queried over World Bank Indicators API. World Bank databank API implements RESTful interface that enables users programmatically access more than 8000 indicators through parameterized queries (Developer Information: Overview, 2017). User initiates a data update macro, selected indicators are returned and recorded into the database (See Appendix 4 for the operational instructions). The querying process can be time consuming and requires high download speed. For instance, if user desires to update data regarding all pre-selected indicators (see chapter 5.1) it can take more than 1 hour to refresh the data. By default, indicator values regarding all countries and time periods are queried (parameterized query for country and time period was tested, but this was discarded as it caused the querying process to become too time-consuming). World Bank data constitutes ca 95% of the data stored in the database, making majority of the database contents updatable. While data was pulled to the database via World Bank API the success of the subsequent pulls is dependent on the stability of the indicator name definitions, which provide basis to the pull.

Figure 5 Solution architecture



Data from other sources is imported to database with MS Excel import. After careful consideration of programmatic options, it was deemed unreasonable, considering the effort required and the benefit it would yield. Firstly, data from other sources constitutes less than 5% of the total data. Secondly, most of the data from these sources is not updated on annual basis and these data sources' data was available only in flat file formats. The files had to be drastically cleaned and modified before data could be imported to the database making it unreasonable to do it programmatically.

Acquired data is first stored in the database in its natural format. Through transformations, data presentation and analysis layers are created (see chapter 4.6 and 4.7). Analysis layer is presented partly in *views*, which will provide flexibility to modify the views. Data is exported to data analysis program either through Excel or over ODBC connection. ODBC driver must be installed in data analysis software before this data import can be performed. Database is shared over suitable sharing platform (Dropbox, Google Drive) with all research team members. See Appendix 4 for the overview of the operating instructions.

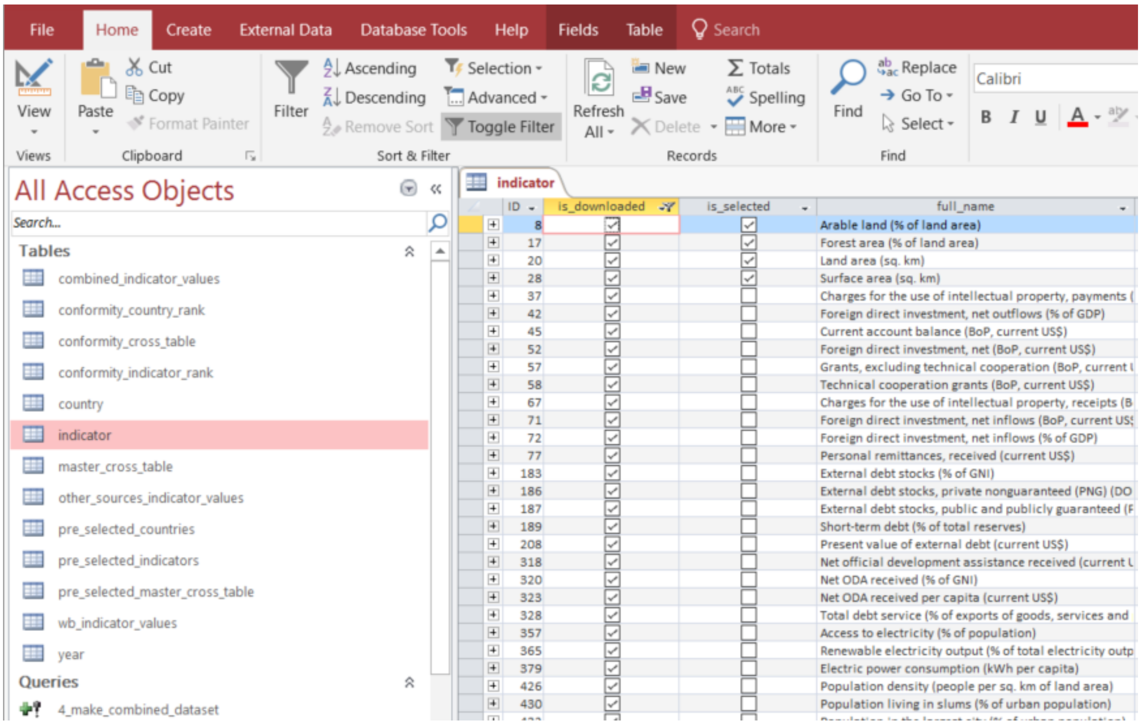
4.5 Database design process

As discussed in chapter 2, an iterative model-based development methodology has been followed to design and build the database. No clear distinction could be drawn between the conceptual, logical and physical design phases (in contrast to the traditional database design process). First version of the conceptual data model was drafted quickly and implemented immediately in MS Access with initial sample data. This initial conceptual data model was used to build a mutual understanding with the research team regarding information requirements and meaning of data. Initial model served as a prototype, which was refined during each following iteration (in cooperation with the research team), based on the following considerations:

- Research domain;
- Main data objects (entities) according which data is collected;
- The natural format of the data;
- Additional information that should be contained in the database (added attributes);
- Desired format of data in output tables and analysis views.

Throughout the process the data model was tested and validated against user's requirements. The aim of the process was to keep the number of tables and columns as minimal as possible, yet as numerous as needed in order to contain all the relevant data. Best overview of the database is obtained by opening the database. Link to the database is given in chapter 6.

Figure 6 Snapshot of the database



Tables
- combined_indicator_values
- conformity_country_rank
- conformity_cross_table
- conformity_indicator_rank
- country
- indicator
- master_cross_table
- other_sources_indicator_values
- pre_selected_countries
- pre_selected_indicators
- pre_selected_master_cross_table
- wb_indicator_values
- year

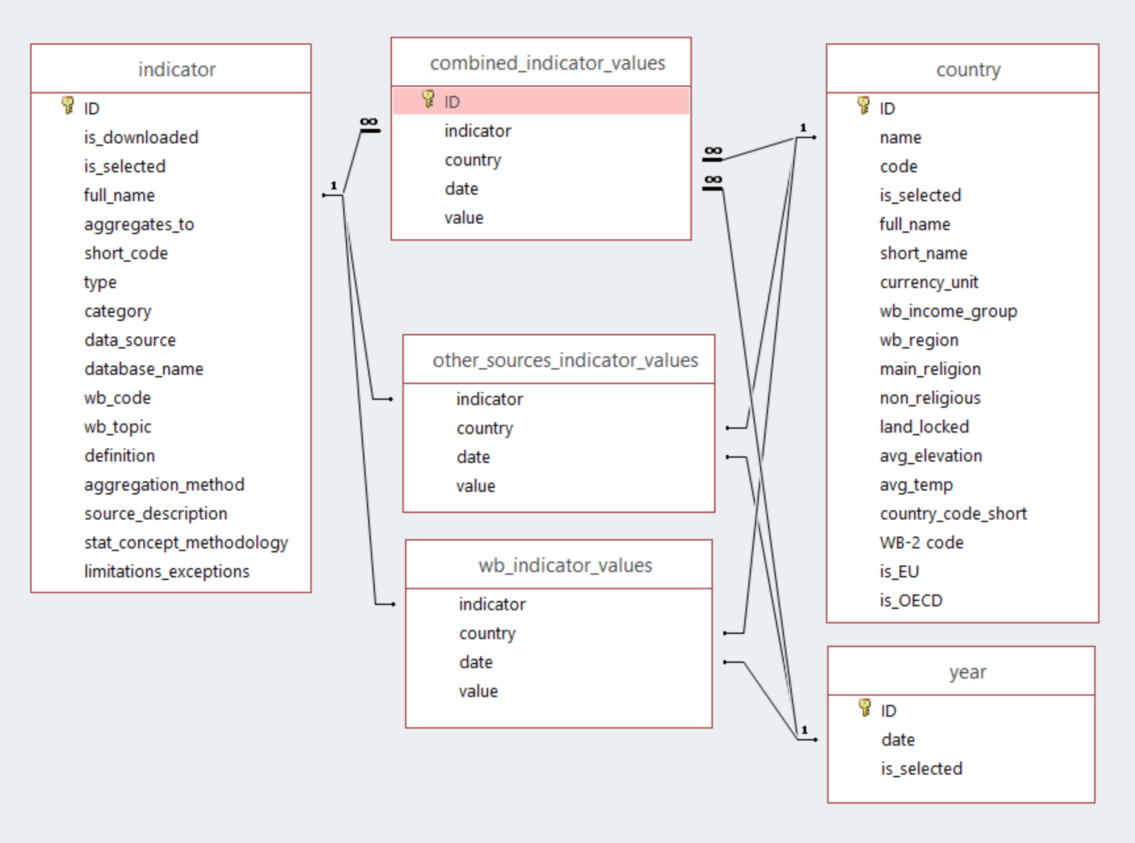
Queries
- 4_make_combined_dataset

ID	is_downloaded	is_selected	full_name
8	☑	☑	Arable land (% of land area)
17	☑	☑	Forest area (% of land area)
20	☑	☑	Land area (sq. km)
28	☑	☑	Surface area (sq. km)
37	☑	□	Charges for the use of intellectual property, payments (
42	☑	□	Foreign direct investment, net outflows (% of GDP)
45	☑	□	Current account balance (BoP, current US$)
52	☑	□	Foreign direct investment, net (BoP, current US$)
57	☑	□	Grants, excluding technical cooperation (BoP, current U
58	☑	□	Technical cooperation grants (BoP, current US$)
67	☑	□	Charges for the use of intellectual property, receipts (B
71	☑	□	Foreign direct investment, net inflows (BoP, current US$
72	☑	□	Foreign direct investment, net inflows (% of GDP)
77	☑	□	Personal remittances, received (current US$)
183	☑	□	External debt stocks (% of GNI)
186	☑	□	External debt stocks, private nonguaranteed (PNG) (DO
187	☑	□	External debt stocks, public and publicly guaranteed (F
189	☑	□	Short-term debt (% of total reserves)
208	☑	□	Present value of external debt (current US$)
318	☑	□	Net official development assistance received (current U
320	☑	□	Net ODA received (% of GNI)
323	☑	□	Net ODA received per capita (current US$)
328	☑	□	Total debt service (% of exports of goods, services and
357	☑	□	Access to electricity (% of population)
365	☑	□	Renewable electricity output (% of total electricity outp
379	☑	□	Electric power consumption (kWh per capita)
426	☑	□	Population density (people per sq. km of land area)
430	☑	□	Population living in slums (% of urban population)

4.6 Data model

The central constructs of the database are *country*, *indicator* (can be viewed as objects/entities/dimensions) and *indicator value* (can be viewed as a fact). As discussed earlier the database does not represent a typical OLTP or OLAP system, but resembles most to a data lake, containing semi-structured and structured data. Hence database is not fully normalized, but it is also not fully unnormalized (corresponds to Second Normal Form, applicable only to base tables). Database tables could be viewed in 2 broad categories (see Table 3): base/source tables and output tables. The data model (see Figure 7) refers to the base tables of the database. Tables *indicator*, *country* and *year* represent semi-structured datasets, which have been formed mostly based on World Bank data. These tables hold time invariant structural information about indicators and countries. These tables will form dimensions for ‘facts’. The central ‘fact’-alike component of the database are tables *wb_indicator_value* and *other_sources_indicator_value*. These two tables hold the indicator values queried from World Bank or imported from Excel in raw formats and through series of queries combined into new table *combined_indicator_values*, which will be used as a key source table to produced desired output tables.

Figure 7 Data model



Tables *combined_indicator_values*, *master_cross_table*, *conformity_cross_table*, *conformity_country_rank* and *conformity_indicator_rank* form the output layer of the database, which are derived as a result of a query (or sequence of queries). The contents of these tables are subject to the parameter selections made in tables *indicator*, *country*, and *year* (See Appendix 1, 2, 3, 4 which give overview of the database elements and provide instructions of operations). Table 3 represents overview of the tables in the database (See Appendix 1 for technical and qualitative table descriptions).

Table 3 Overview of database tables

Table name	**Table type**	**Data source**	**Macro**
indicator	source	Import from Excel	
country	source	Import from Excel	
year	source	Import from Excel	
other_sources_indicator_values	source	Import from Excel	

Table name	Table type	Data source	Macro
wb_indicator_values	source	Automatic over World Bank API (initiated with a macro)	1 Update World Bank Data
combined_indicator_values	source/ output	Result of a query (initiated with a macro)	2 Update data selection
master_cross_table	output	Result of a query (initiated with a macro)	3 Update master cross table
conformity_cross_table	output	Result of a query (initiated with a macro)	6 Update conformity cross table
conformity_country_rank	output	Result of a query (initiated with a macro)	4 Update conformity analysis countries
conformity_indicator_rank	output	Result of a query (initiated with a macro)	5 Update conformity analysis indicators

In addition to the tables above certain other tables have been saved as permanent data tables. These tables indicate the pre-selection of indicators (*pre_selected_indicators*) and countries (*pre_selected_countries*) performed by the author during this thesis and the output table (*pre_selection_master_cross_table*) formed based on this pre-selection (see chapter 5.1). User of the database can restore the author pre-selection with respective queries saved in the database (See Appendix 4).

Data types

In all source tables data types have been selected considering the possible values in the fields that correspond to the columns and what operations must be performed with these columns. Wherever possible data types have been chosen to optimize the database capacity (See Appendix 1).

Referential integrity

Referential integrity has been enforced among the source tables *country*, *indicator*, *year* and *combined_indicator_values*, as these tables form the basis for all the output tables (through queries). Since *wb_indicator_values* and *other_sources_indicator_values* are not controlled by the system then referential integrity cannot be forced on that level, but it is done in com*bined_indicator_values* table, which combines data in cleaned format from both sources tables. Undefined relationships (See Figure 7) have been assigned for tables *wb_indicator_values* and *other_sources_indicator_values.*

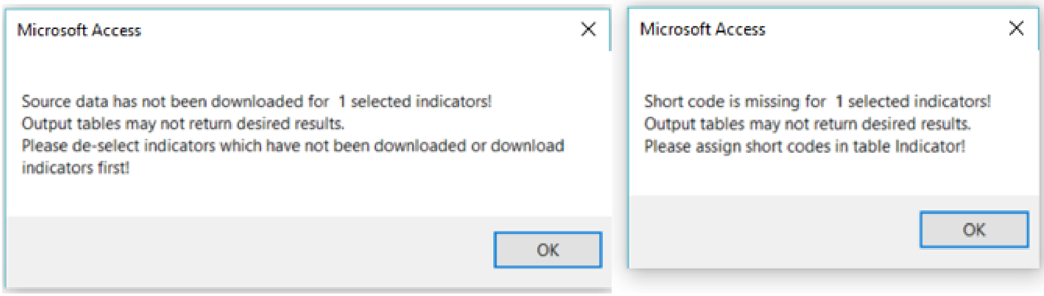
Primary and foreign keys

All primary keys are surrogate keys (See Figure 7). Generally, it is not advisable to use surrogate keys (Eesaar, 2008), but in this case surrogate keys have been implemented to speed up indexing and also to provide overview of the number of instances in the table. All foreign keys are alternate keys, which have been enforced through constraints (fields are required, no duplicates, indexed). See Appendix 1 for technical source table overviews.

Constraints

Necessary constraints have been implemented through referential integrity, data types and value constraints. While not critical for running the queries and macros, certain attribute values need to correspond to certain logic in order to produce output tables with desired information and in desired format. MS Access 2016 does not allow to build validation rules based on two different columns, hence violation of the rule is delivered through a message box to the user. Rule itself is checked with SQL and VBA. One of such rules relates to the selection of indicators. Only such indicators should be selected (‘is_selected’=Yes) that have been downloaded (‘is_downloaded’=Yes) previously to the database. Additionally, all selected indicators should have assigned short codes. Short codes are important as in some output tables each short code becomes column heading (standard panel format). In order to update the selection of indicators and to be able to produce output tables based on the new selections, user needs to trigger macro *2 Update data selection*. In the end of this macro user is informed if any of these rules were violated (see Figure 8) and gives instructions what to do.

Figure 8 Rule violation messages



Source data has not been downloaded for 1 selected indicators!
Output tables may not return desired results.
Please de-select indicators which have not been downloaded or download indicators first!

Short code is missing for 1 selected indicators!
Output tables may not return desired results.
Please assign short codes in table Indicator!

In addition to permanent data tables, the database also includes other objects (see Figure 6 Snapshot of the database and Appendix 1, 2 and 3):

- Queries (27), which are necessary for data transformations and presenting views (virtual tables);
- Macros (6), tools which contain commands to automate the data update process in the output tables;
- Modules (6), objects (set of functions, variables and routines written in VBA code) which are used in macros.

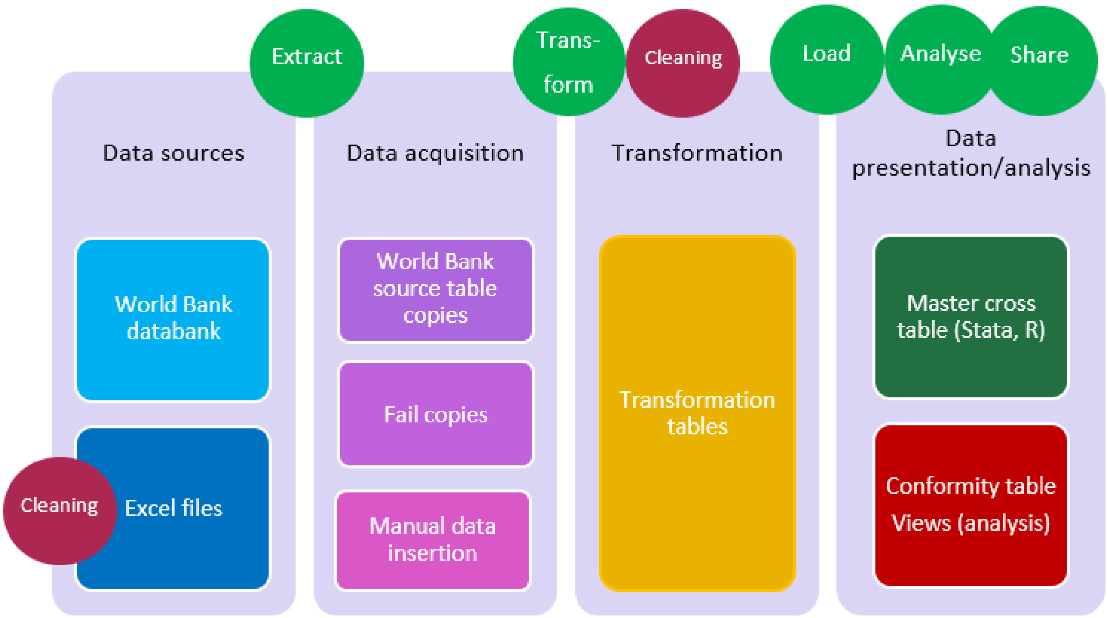
4.7 ETL processes

Next, the data flow from source layers to output layers will be described. As discussed earlier, then the database designed and built during this thesis is a custom built “tool” to facilitate the data pre-processing, -analysis and feature selection process for the research team. Hence, the data flow process does not follow a standard ETL process.

The term “ETL” is widely used as a broad term referring to data extraction from the source system(s) and subsequent loading into the warehouse. ETL stands for “extraction”, “transformation” and “loading. Although often viewed as three distinct steps, the process is rarely such and includes also “transportation” step, during which data is physically transported to the target system (Oracle, 2017). The main purpose of ETL process is to facilitate the process of data analysis and reporting by ensuring the data is readily usable in standardized and validated form.

Although ETL process necessary for this database is much simpler and less complex than for data warehouses that are supporting business environments, the goal of the process is the same: to produce structured and clean data in required formats. The figure below illustrates the layers of the database and the data flow:

Figure 8 Layers of the database



The database is divided into layers based on the conceptual purpose of the object rather than object type (tables, macros, queries, modules).

Extraction

First, data is extracted from the data sources. The main data source is World Bank databank. To query data from World Bank a selection of indicators must be done previously in table *indicator*. Data extraction is initiated with macro *1 Update World Bank data* (see module code and queries in Appendix 2 and 3). Data is extracted over World Bank API and stored in its raw format in table *wb_indicator_values*. Data, which is previously rendered into unanimous format from other sources is imported from Excel tables and stored in table *other_sources_indicator_values*. The Excel files have been cleaned and formatted to match the data types and formats of the *wb_indicator_values* table. Some columns for selected indicators must be filled manually before data can be transformed into output tables. As the short codes will become column headings in final data output table, STATA requirements for column names were taken into consideration (only letters, digits and underscores can be used (StataCorp LP, 2013)), max of 8 characters were used to keep them short). Type refers to the indicator measurement scale and category to the category under which it was classified (classification given in Appendix 1). Extraction is not an ongoing process – it is initiated manually only when an update to the data is desired (for example, if new data points have become available). Data stored in the database is on annual basis, hence the data update should not be needed very frequently (in case other indicator than those preselected and downloaded or if new year data is published by the data source). See Appendices 1 - 4 that give overview of the database elements.

Transformation and loading

Next, data from two source tables is merged into one cohesive dataset into table *combined_indicator_values* with macro *2 Update data selection* (see module code in Appendix 3). During this transformation three additional tables are being created and later deleted. Alternatively, the same process can be carried out by triggering individual queries (numbered from 1 to 4), in which case the process is more easily controlled and in case of a failure easier to troubleshoot. These temporary tables, where the foreign keys are used to decrease data size, are necessary for type (for column 'value') conversions (changing datatype from short text as given by World Bank to double to enable calculations with the data). During type conversion the database size is expanding exponentially. Thus, in order for the process to succeed maximum of 154 countries and 231 World Bank indicators can be selected over 30-year period at a time (this is the maximum limit tested which was successful). While this poses a restriction to the task, it is more than unlikely that more than 154 countries, 231 indicators and 30 years are selected for creating output tables and analysis (confirmed with the research team). In a likely case, database user is interested in investigating indicators of one category (depending on the category, one category includes 10 – 80 indicators) over 10-20 years. Hence the probability of surpassing the maximum data selection is highly unlikely. In future, this could be further optimized by splitting the databases into back/ and frontend bases.

Output table *combined_indicator_values* will hold data regarding indicators, countries and years that were selected (column ‘is_selected’ is ticked off) by the user; the selection must be implemented in this phase as otherwise the output tables would become too large to be saved as tables. Next, the data in table *combined_indicator_values* is transformed into panel data format by initiating macro *3 Update master cross table* and result is saved into *master_master_cross_table* (See Appendix 1, 2, 3 and Appendix 4 for operating instructions).

Cleaning

Very little cleaning is needed for data which is pulled from World Bank. However, substantial amount of data cleaning is required for data that will be loaded from Excel files. All data sources provide data in different formats and it was not considered reasonable to automate the cleaning process for these sources. Thus, all data originating from other sources was cleaned and transformed manually in Excel to render it to a format that can be merged with World Bank data.

A lot of data gathered is of qualitative nature. In these cases, numerical values have been assigned to these categorical variables in order to record the data as qualitative (but these values have no quantitative significance).

Analysing

Different views represent the analysis layer of the database. Various queries (see Appendix 2 and Appendix 4) are used to provide initial overview of the indicators (descriptive statistics), indicators’ data quality and availability across selected indicators and countries (see chapter 5). With macros 4-5 it is possible to run conformity analysis on the selected data. Views and results of the conformity analysis can be used to modify the indicator, country and year selection and produce new versions of the *master_cross_table*. All query results have been carefully validated either by triggering all sub-queries one by one and randomly checking the results against the data or by comparing the query results against an alternative query.

Sharing

There are two options to export data to desired data analysis program. First and perhaps easier option is to simply export desired table/query into Excel file (External

Data=>Export Excel) and then import the file again into data analysis program. Second option is to set up an ODBC connection, but this can be done only if MS Access driver has been set up and the versions of the MS Office (MS Access) and data analysis programs match. For instance, user needs to have installed 64bit Office and 64bit STATA for this option to work. See Appendix 4 that shows to import data to Stata over ODBC.

5 Data pre-analysis

The aim of the pre-analysis phase is to explain the process of selecting initial subset of indicators as well as design and implementation of tools that would enable comparative analysis of the indicator usability for further econometric modelling.

The main components of the pre-analysis are as follows:

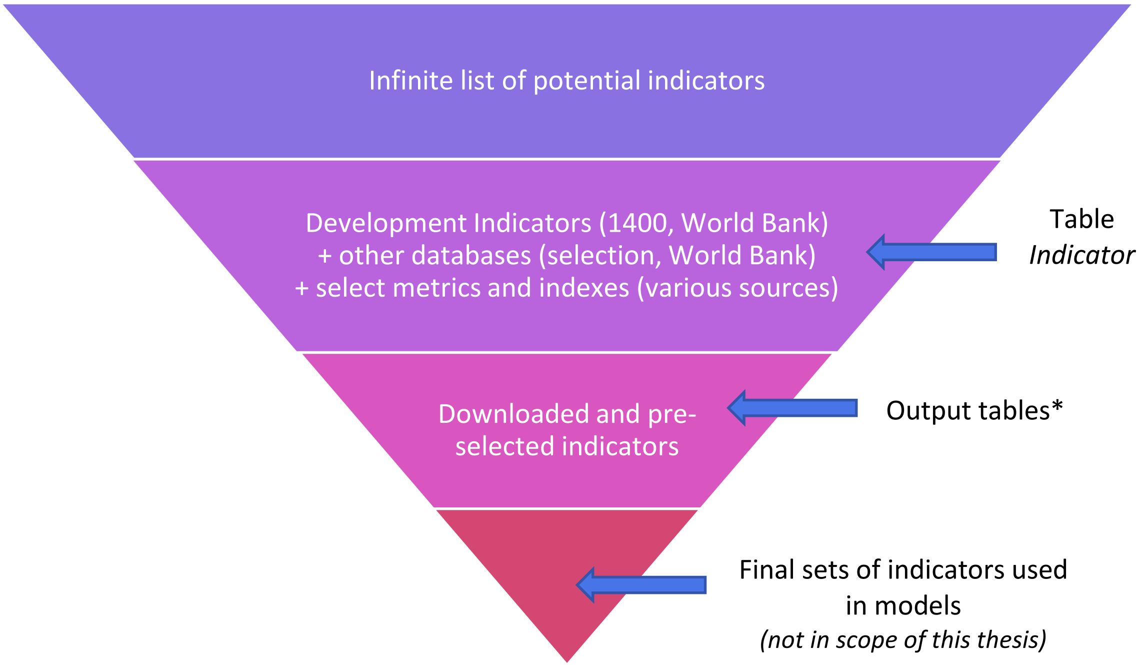
- Description of the initial feature selection;
- Construction of views with SQL for comparative descriptive statistics;
- Overview of the implementation of conformity analysis with SQL.

These steps should facilitate researchers to identify subsets of indicators across different countries and domains with best statistical strength, quality and availability, which should in turn decrease the number of iterations in economic modelling process and avoid time-consuming data mining activities occurring in econometric modelling phases.

5.1 Feature selection

Feature selection is a method of data mining used in preliminary stages of research, where out of large list of candidate variables a manageable subset of variables is chosen for further analysis (StatSoft, 2013). Such approach is very common when data is collected via (partially) automated methods. The feature selection is based on thorough research into knowledge indicators (see chapter 3) and relevant data sources. The measurement of KBD is a complex topic (see chapter 3) and there are hundreds of potentially useful indicators to choose from. Furthermore, the research project is interested in many subsections of KBD. Thus, it was not possible and reasonable to identify the specific subset (of indicators) of data immediately and more data than actually necessary was initially extracted. The final subset of indicators was identified through several iterations as illustrated below.

Figure 9 Feature selection process


Infinite list of potential indicators
Development Indicators (1400, World Bank)
+ other databases (selection, World Bank)
+ select metrics and indexes (various sources)
Table
Indicator
Downloaded and pre-selected indicators
Output tables*
Final sets of indicators used in models
(not in scope of this thesis)

** Output tables refer to the tables which are generated based on the user selections. Author's indicator and country pre-selections are saved as separate tables pre_selected_indicators, pre_selected_countries, pre_selection_master_cross_table.*

List of indicators

Indicators (full indicator list in table *indicator*) contained in the database represent a 'long list' of indicators potentially useful for the research project (see Table 4). The potential data sources and indicators were sought after based on the data requirements identified in chapter 2. Author's focus was on finding reliable indicators and metrics with global coverage which would characterise the institutional regime and efficiency, various dimensions of governance, quality of business environment (including efficiency of and access to capital and labour markets, investment climate and ease of doing business) and other critical knowledge creation, absorption and diffusion measures.

The main data source is World Bank with its sub-databases. World Development Indicators (WDI), presenting a comprehensive list of indicators (1400), useful for assessing a country's general development level, form the core of the indicator list. Such broad coverage of key indicators across various sectors is useful as it is likely to offer several alternatives for each domain. World Bank was selected as the main data source, since it has by far the best and most comprehensive set of data across all countries. Additionally, World Bank is the only international institution that offers API connection to its data. Most of the renowned international institutions still share their data via flat files. Other major international institutions such as OECD, IMF and Eurostat etc. were also explored for data, yet discarded since either their data is already included in World Bank databank or they fail to provide required geographic coverage.

Although WDI database offers surprisingly good coverage to the data requirements identified, it was insufficient for covering all data requirements. Many other interesting potentially useful indicators and data sources were discovered and included into the list of indicators. In addition to the WDI database indicators, database includes more than 270 indicators measuring economic and institutional regime and the development towards economic freedom sourced from World Bank database (Country Policy and Institutional Assessment, Doing Business, Enterprise Surveys, World Governance Indicators) as well as other reputable institutions such as Bertelsmann Institution, Freedom House, Fraser Institute, Reporters Without Borders and Swiss Economic Institute.

Table 4 Overview of indicators included in the database

Main data source	**Sub-database (if applicable)**	**Domain/description**	**Included in database**	**Period covered**	**Nr of countries covered***
World Bank	World Development Indicators	World Bank primary collection of development indicators across wide array of topics (agriculture, economy & growth, education, energy, environment, financial Sector, health, infrastructure, private sector, public sector, science & technology, etc.) (1400 indicators)	All	1960-2016	80-150 Depending on the specific indicator
World Bank	Education Statistics	Collection of internationally comparable indicators describing education access, progression, completion, literacy, teachers, population, and	Selection of key indicators covering literacy rates, government expenditure of	1970 - 2100	60 -70 Depending on the specific indicator

Main data source	**Sub-database (if applicable)**	**Domain/description**	**Included in database**	**Period covered**	**Nr of countries covered***
		expenditures. The indicators cover the education cycle from pre-primary to vocational and tertiary education. (4000 indicators)	education, secondary school attendance rates and PISA test results. (30 indicators)		
World Bank	Country Policy and Institutional Assessment (CPIA)	Rating of countries against a set of 16 criteria grouped in four clusters: economic management, structural policies, policies for social inclusion and equity, and public sector management and institutions. (21 indicators)	All	2005-2014	63
World Bank	Doing Business	Measures of business regulations and their enforcement. (58 indicators)	All	2004-2016	135-147 Depending on the specific indicator
World Bank	Enterprise Surveys	Firm-level data from over 125,000 establishments in 139 countries. Data are used to create over 100 indicators that benchmark the quality of the business environment across the globe. Each country is surveyed every 3 to 4 years. (121 indicators)	All	2005-2014	100-118 Depending on the specific indicator
World Bank	World Governance Indicators	Worldwide Governance Indicators capture six key dimensions of governance (Voice & Accountability, Political Stability and Lack of Violence, Government Effectiveness, Regulatory Quality, Rule of Law, and Control of Corruption) (6 indicators)	All	1996-2016	149
Bertelsmann Foundation	Status Index, Management Index	Indicators measuring how developing countries are steering social change toward democracy and a market economy.	2 key indexes + 23 component indicators	2006-2016	118
Fraser Institute	Economic Freedom Summary Index	Index measures the degree of economic freedom present in five major areas (Size of government, legal system and security of property rights; sound money; freedom to trade internationally, regulation.)	1 key index + 36 component indicators	1970-2014	140
Freedom House	Freedom Status Rating	Measures the degree of civil liberties and political rights.	3 key indicators	1972-2016	150

Main data source	Sub-database (if applicable)	Domain/description	Included in database	Period covered	Nr of countries covered*
Reporters Without Boarders	World Press Freedom Index	Measures the degree of freedom available to journalists in 180 countries.	1 key indicator	2002-2017	151
KOF Swiss Economic Institute	KOF Index of Globalization	Measures the economic, social and political dimensions of globalization.	1 key indicator	1970-2013	148
Transparency International	Corruption Perception Index	Measures perceived levels of corruption, as determined by expert assessments and opinion surveys.	1 key indicator	2012-2016	150
CIA (Factbook)	Former and current socialist states	Indicates the former and current socialist states.	1 key indicator	1970-2017	151

** Out of the 154 selected countries*

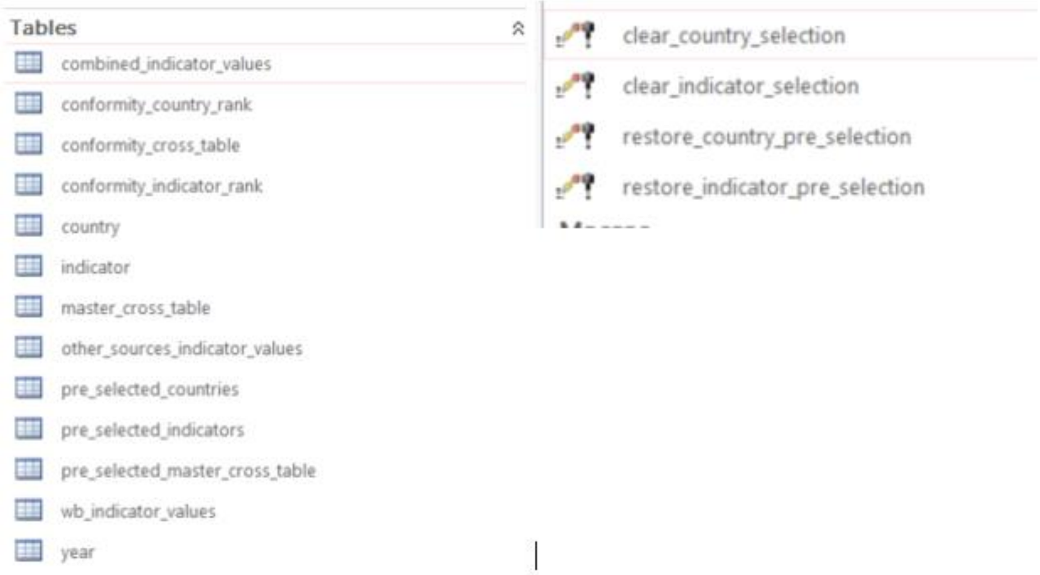
Pre-selected indicators (feature selection)

Pre-selected indicators represent a sizable pool of metrics, of which further suitability into econometric models can be now assessed through conformity analysis and comparative analysis of indicator parameters such as the quality, availability and dispersion. Author has conducted initial selection of 231 indicators (see table *pre_selected_indicators,* see Appendix 4) from the data available in the database considering the economic theory and background of the research as well as availability and relevance of the indicators in the database. The pre-selection contains 6+7 structural, 11 demographical, 31 education system and level, 11 health, 82 economic performance, 12 innovation system, 6 ICT infrastructure, 47 business environment and 28 institutional regime and efficiency related indicators. Many indicators could be classified under more than one category; therefore, the classification is tentative. Full list of pre-selected indicators is attached in Appendix 5. The pre-selection forms the second iteration in feature selection process.

The final output table (*master_cross_table*) is subject to some other data selections such as the selection of countries and time period. Out of 217 countries contained in the database, 154 major economies were selected (see table *pre_selected_countries*) discarding small and insignificant (island) economies. While World Bank data stretches back to 1960’s (but is rather limited), data from other sources is fairly limited before the 2000’s, hence time period from 1980 – 2017 has been selected (in table *year* column ‘is_selected’ is ticked off). All these pre-selections can be modified (by using information obtained from the comparative overview of indicators (view1, view2 and conformity analysis – see chapter 5.2) to produce customized sets of data, which will be used for econometric modelling.

The pre-selections have been recorded as permanent data tables (see Figure 10) to provide evidence to the pre-selection. Database is delivered with random selection of indicators and countries, users can restore the selection proposed by the author by saved queries (see Figure 10 to the right). See Appendix 4 for instructions.

Figure 10 Pre-selection tables and queries for restoring the pre-selection



Tables

- combined_indicator_values
- conformity_country_rank
- conformity_cross_table
- conformity_indicator_rank
- country
- indicator
- master_cross_table
- other_sources_indicator_values
- pre_selected_countries
- pre_selected_indicators
- pre_selected_master_cross_table
- wb_indicator_values
- year

- clear_country_selection
- clear_indicator_selection
- restore_country_pre_selection
- restore_indicator_pre_selection

Final sets of indicators

Assembling final sets of indicators used in econometric models are out of the scope of this thesis. The final selection is subject to the outcomes of this thesis and is made by the research team after having carefully studied the availability, quality and statistics made available during this thesis. No meaningful statistical method can be used on 231 indicators. Along with the growth of dimensionality, the amount of data needed to produce statistically reliable results grows exponentially (StatSoft, 2013). Feature selection process is usually iterative (StatSoft, 2013) in its nature and will be performed for each economic model, yielding in a specific subset of indicators for each model.

5.2 Design and implementation of pre-analysis tools

Once the pre-selection of indicators, countries and years is done, one naturally wishes to ‘see’ how the data looks like. It is hard to get overview of the quality of data from the standard panel view when it involves thousands of lines and hundreds of columns. Two types of tools have been created to give fast and simple overview of the selected data and to help to refine the selection:

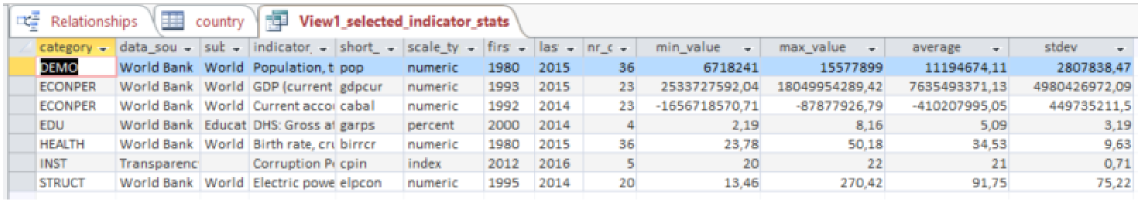
1. Views based on SQL queries presenting various parameters of the indicators. SQL queries are easily adjustable if needed.

2. Conformity analysis, implemented with SQL and VBA.

5.2.1 Indicator descriptive statistics

In order to simplify the large amount of data collected and establish initial overview of the selected data and allow for meaningful comparison of the indicator data collected on country basis, some simple descriptive statistics are calculated and presented along with other potentially useful parameters (see Figure 11 and Appendix 5) with the help of parameterized query (*view1_selected_indicator_stats*) (see Appendix 2). This view could be useful either on group of countries or on one country (see example below). The countries could easily be swapped/added by modifying the query code (*Design view=>SQL view*). Immediate results can be saved in temporary tables and/or in duplicate queries and compared to each other.

Figure 11 Snapshot of view View1_selected_indicator_stats



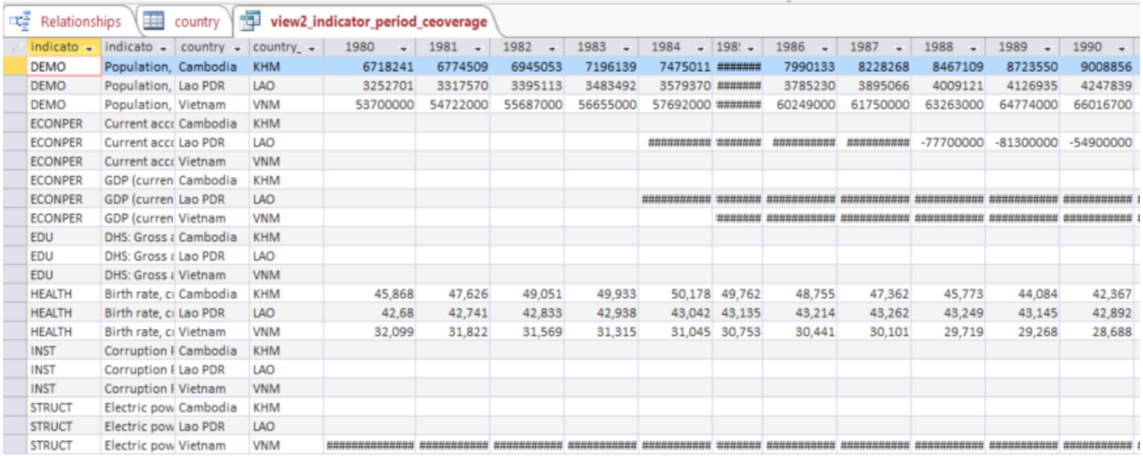
category	data_sou	sub	indicator	short_	scale_ty	firs	las	nr_c	min_value	max_value	average	stdev
DEMO	World Bank	World	Population, t	pop	numeric	1980	2015	36	6718241	15577899	11194674,11	2807838,47
ECONPER	World Bank	World	GDP (current	gdpcur	numeric	1993	2015	23	2533727592,04	18049954289,42	7635493371,13	4980426972,09
ECONPER	World Bank	World	Current acco	cabal	numeric	1992	2014	23	-1656718570,71	-87877926,79	-410207995,05	449735211,5
EDU	World Bank	Educat	DHS: Gross at	garps	percent	2000	2014	4	2,19	8,16	5,09	3,19
HEALTH	World Bank	World	Birth rate, cru	birrcr	numeric	1980	2015	36	23,78	50,18	34,53	9,63
INST	Transparenc		Corruption Pe	cpin	index	2012	2016	5	20	22	21	0,71
STRUCT	World Bank	World	Electric powe	elpcon	numeric	1995	2014	20	13,46	270,42	91,75	75,22

Select descriptive statistics include: number of observations, first year of observation, last year of observation, minimum value, maximum value, average, and standard deviation. Measures of central tendency (average/mean) and variability (standard deviation, minimum, maximum values) help to understand the nature of the data. In addition to basic descriptive statistics, number of observations along with the first and last year when this indicator is available have been presented for each indicator. Indicators with higher observation count should be preferred. Furthermore, the measurement scale type and category of each indicator is presented (see Appendix 1 for database table descriptions and scale type and category definitions). Some statistical analysis is only meaningful for data measured at certain scales. All this additional information (observation count, dispersion, central tendency, scales, availability and temporal continuity) regarding the indicators have an impact on the assessment of the "statistical strength" of indicators and on selecting suitable statistical method and/or economic model.

5.2.2 Indicator availability

Although previous view (*view1_selected_indicator_stats*) did provide information regarding indicator general availability, quality and consistency on select country level, this information was not in the best format to gauge it visually. Thus, an additional overview (see Figure 12) of the indicator availability across time was constructed using a parameterized query *view2_indicator_period_coverage* (see Appendix 5 and Appendix 2), which pivots the data so that each year becomes an attribute. In the example presented below Vietnam, Cambodia and Laos and select institutional indicators are selected and placed on the timeline.

Figure 12 Snapshot of view View2_indicator_period_coverage



indicato	indicato	country	country_	1980	1981	1982	1983	1984	198!	1986	1987	1988	1989	1990
DEMO	Population,	Cambodia	KHM	6718241	6774509	6945053	7196139	7475011	#######	7990133	8228268	8467109	8723550	9008856
DEMO	Population,	Lao PDR	LAO	3252701	3317570	3395113	3483492	3579370	#######	3785230	3895066	4009121	4126935	4247839
DEMO	Population,	Vietnam	VNM	53700000	54722000	55687000	56655000	57692000	#######	60249000	61750000	63263000	64774000	66016700
ECONPER	Current acc	Cambodia	KHM											
ECONPER	Current acc	Lao PDR	LAO					##########	#######	##########	##########	-77700000	-81300000	-54900000
ECONPER	Current acc	Vietnam	VNM											
ECONPER	GDP (curren	Cambodia	KHM											
ECONPER	GDP (curren	Lao PDR	LAO					###########	#######	###########	###########	###########	###########	###########
ECONPER	GDP (curren	Vietnam	VNM						#######	###########	###########	###########	###########	###########
EDU	DHS: Gross	Cambodia	KHM											
EDU	DHS: Gross	Lao PDR	LAO											
EDU	DHS: Gross	Vietnam	VNM											
HEALTH	Birth rate, c	Cambodia	KHM	45,868	47,626	49,051	49,933	50,178	49,762	48,755	47,362	45,773	44,084	42,367
HEALTH	Birth rate, c	Lao PDR	LAO	42,68	42,741	42,833	42,938	43,042	43,135	43,214	43,262	43,249	43,145	42,892
HEALTH	Birth rate, c	Vietnam	VNM	32,099	31,822	31,569	31,315	31,045	30,753	30,441	30,101	29,719	29,268	28,688
INST	Corruption I	Cambodia	KHM											
INST	Corruption I	Lao PDR	LAO											
INST	Corruption I	Vietnam	VNM											
STRUCT	Electric pow	Cambodia	KHM											
STRUCT	Electric pow	Lao PDR	LAO											
STRUCT	Electric pow	Vietnam	VNM	##############	###########	###########	###########	###########	#######	###########	###########	###########	###########	###########

Such presentation helps to understand better the indicator availability across categories and countries. In the example above, significant differences in the availability of data across these three countries can be observed. As such, it can be concluded that it is not possible to run any time-series method across these three countries regarding selected indicators as there are very few observations for Cambodia and Lao available. Therefore, alternative indicators must be sought after, or some other statistical method must be used (some multivariate method). Additionally, the view is also useful in determining the time slots for the models and assessing the need for data imputation.

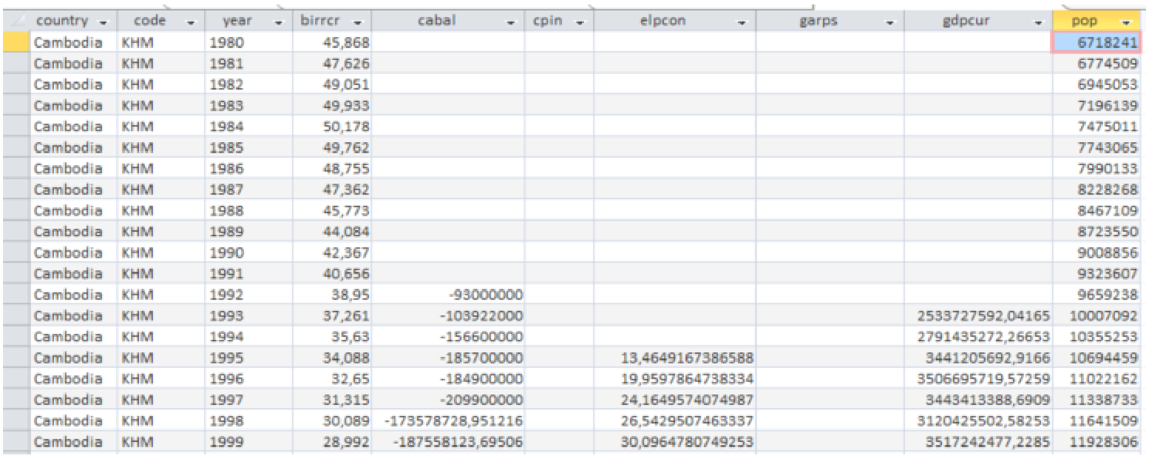
5.2.3 Conformity analysis

Last step of the pre-analysis phase is implementation of conformity analysis with MS Access 2016 SQL. Conformity analysis is a data mining method based on the Monotone Systems Theory, developed by group of researchers at the Tallinn University of Technology, in which main goal is to reorganize data according to specific property – conformity, which is essentially a measure of frequency. Conformity analysis is an alternative to classification and clustering; it aligns the objects and attributes according to nearest-neighbour similarity and therefore establishes a scale of typicality in the data (Liiv, Kuusik, & Võhandu, 2007). During conformity analysis N*M data matrix will be reorganized based on the ranking of elements in rows and columns, which will allow to visually discover patterns in the data (clusters) and easily detect "typical and fuzzy parts of the data" (Kuusik, Lind, Võhandu, 2004).

Conformity analysis is especially useful in the context of this thesis and research. One of the sub-goals of this thesis is to implement tools which will allow the researchers to determine countries (those that will accompany the target countries Vietnam, Cambodia and Lao) and indicators (from wider pool of indicators) with best data quality. Conformity analysis will help to achieve this goal by reorganizing the data matrix (those countries, years and indicators which are ticked off in column 'is_selected' in respective source tables) so that the countries (left axis) and indicators (right axis) with the best 'conformity' appear at the left-most corner of the table (see Figure 13 and 14). Since most of the attributes contained in the database are not categorical, the standard approach of frequency measure is modified to indicate the temporal frequency (value exists=1 else 0). Additionally, the approach is adapted to address the three-dimensionality of the data (country, indicator, year) versus the standard two-dimensional approach. Frequencies are calculated across countries, indicators and years. Another modification relates to the process of ranking. In instances when there are multiples countries or indicators with equal scores, the top object is selected from table (no further metrics are calculated to decide the highest ranking).

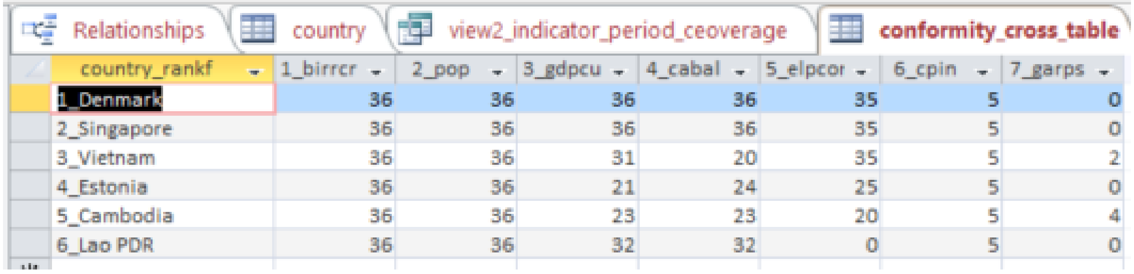
Conformity is a measure of relative frequency and the values represent the count of yearly observations (e.g. indicator *birrcr* is observable over 36 periods for Denmark out of 37 periods). *Birrcr* is most conformant indicator, Denmark is most conformant country. For instance, if there were two measures of GDP with different conformity scores, then the one with higher score (all else equal) would be better choice into the final output table. Such representation helps to determine set of countries and indicators with best data quality. In econometric analysis the length of time lines is very critical, hence the frequency of data is a critical measure.

Figure 13 Snapshot from initial unordered dataset (example with 6 countries and 7 indicators)



country	code	year	birrcr	cabal	cpin	elpcon	garps	gdpcur	pop
Cambodia	KHM	1980	45,868						6718241
Cambodia	KHM	1981	47,626						6774509
Cambodia	KHM	1982	49,051						6945053
Cambodia	KHM	1983	49,933						7196139
Cambodia	KHM	1984	50,178						7475011
Cambodia	KHM	1985	49,762						7743065
Cambodia	KHM	1986	48,755						7990133
Cambodia	KHM	1987	47,362						8228268
Cambodia	KHM	1988	45,773						8467109
Cambodia	KHM	1989	44,084						8723550
Cambodia	KHM	1990	42,367						9008856
Cambodia	KHM	1991	40,656						9323607
Cambodia	KHM	1992	38,95	-93000000					9659238
Cambodia	KHM	1993	37,261	-103922000				2533727592,04165	10007092
Cambodia	KHM	1994	35,63	-156600000				2791435272,26653	10355253
Cambodia	KHM	1995	34,088	-185700000		13,4649167386588		3441205692,9166	10694459
Cambodia	KHM	1996	32,65	-184900000		19,9597864738334		3506695719,57259	11022162
Cambodia	KHM	1997	31,315	-209900000		24,1649574074987		3443413388,6909	11338733
Cambodia	KHM	1998	30,089	-173578728,951216		26,5429507463337		3120425502,58253	11641509
Cambodia	KHM	1999	28,992	-187558123,69506		30,0964780749253		3517242477,2285	11928306

Figure 14 Conformity analysis (example with 6 countries and 7 indicators)



country_rankf	1_birrcr	2_pop	3_gdpcu	4_cabal	5_elpcor	6_cpin	7_garps
1_Denmark	36	36	36	36	35	5	0
2_Singapore	36	36	36	36	35	5	0
3_Vietnam	36	36	31	20	35	5	2
4_Estonia	36	36	21	24	25	5	0
5_Cambodia	36	36	23	23	20	5	4
6_Lao PDR	36	36	32	32	0	5	0

5.2.3.1 Implementation in SQL

The conformity analysis has been implemented through MS Access 2016 SQL and MS Access 2016 VBA. It would be rather difficult and very cumbersome to run this analysis in non-automated manner, especially on large data matrices. By using SQL and VBA all calculations steps have been delegated to database system MS Access, allowing for fast and repetitive calculations.

Conformity analysis comprises two sets of iterations during which ranking of countries (macro 4 Update conformity analysis countries) and indicators (macro 5 Update conformity analysis indicators) is determined. The materials of Võhandu et al (2006) and Liiv et al (2007) have been used as a basis to construct the algorithm in SQL (algorithm itself as well as implementation in SQL). Some modifications had to be implemented as data used for this analysis is three dimensional (country, indicator, year) and not two-dimensional, as presented in the materials of Võhandu et al (2006) and Liiv et al (2007).

Based on Võhandu (1989) and Võhandu et al (2006) there are three key methods of reordering data: minus technique, plus technique and mixed technique. Minus technique, based on which country or indicator with the lowest level of 'conformity' has been eliminated from the initial dataset, has been used in this thesis to reorder data.

The algorithm has been implemented through a sequence of queries invoked by the VBA module "Conformity" (see module code and queries used in the code in Appendix 2 and 3). The key steps of the algorithm are following:

1. Ranking of countries (macro 4)
    a. Counting number of countries with observations within indicator and year;
    b. Replacing indicator values with the frequency of observations within that indicator and year;
    c. Calculating conformity of countries as sum of indicator values (count of observations);
    d. Saving country with the smallest indicator count sum into separate table (*conformity_country_rank*);
    e. Eliminating country with the smallest indicator count sum from the initial dataset;
    f. Repeating steps a-e until no rows remain in the initial dataset.
2. Ranking of indicators (macro 5)
    a. Counting number of indicators with observations within country and year;
    b. Replacing indicator values with the frequency of observations within that country and year;
    c. Calculating conformity of indicators as sum of country values (count of observations);

   d. Saving indicator with the smallest country count sum into separate table (*conformity_indicator_rank*);

   e. Eliminating indicator with the smallest country count sum from the initial dataset;

   f. Repeating steps a-e until no rows remain in the initial dataset.

3. Combining and reordering countries and indicators into final data table (*conformity_cross_table*) based on the recorded ranking, initiated with macro 6.

Author chose to implement the algorithm through sequence of queries (for each step of the algorithm there is a separate query) rather than in one or few long queries in order to improve transparency and traceability of the algorithm steps (see Appendix 2 and 3). In this way it was easier to test the result of the queries in each step. It also makes it easier to track and understand each part of the algorithm and when necessary adjust it. The algorithm has been built so that it is fully scalable; number of years, countries and indicators can be changed (by modifying the columns 'is_selected' in tables *country, indicator* and *year,* see also instructions in Appendix 4). Output tables (*conformity_country_rank, conformity_indicator_rank, conformity_cross_table*) will adjust to data additions and reductions. However, it is not advisable to use this algorithm on huge data matrices (above 20x20 data matrices) as the processing may take up very long time due to MS Access data limitation of 2 GB. The resulting table is also too large to visually gauge patterns in the data. As further optimisation (which was not implemented as part of this thesis), each iteration of the conformity method could be initiated by a separate MS Access database, which would also force the other database with the actual conformity calculations to be compacted at each iteration. In the context of data pre-analysis conformity analysis should be run already refined selection of countries and indicators.

All queries that are part of the algorithm have been carefully validated. Results of the steps have been validated against alternative query and/or results obtained by manually running through the iterations in Excel. The queries could be further optimized for speed and memory in the future.

The steps of running the conformity analysis have been described in Appendix 4.

6 Delivery

While the study can be regarded as client-contractor relationship then given high level of interconnectedness between these roles the study did not follow the usual framework or procedures specific for this type of relationship. Database is delivered with random (small) selection of downloaded indicators and countries. Pre-selection of indicators and countries performed during this thesis are saved as permanent data tables and these selections can be restored with respective queries (See Appendix 4) should the user desire to do so.

The deliverables of the study were handed over several phases and included following:

- Database implemented on MS Access platform, that currently contains information about 231 pre-selected indicators;
- Instructions that give simple overview of the database and designed functionality (see Appendix 4);
- Documentation about the main source tables (*country*, *indicator*) and their attributes (see Appendix 1);
- Package of key source and output tables in Excel (as a backup version).

All deliverables have been shared via Google Drive and are accessible on the following link:

https://drive.google.com/open?id=1I2-lpt0mo3vhwgJLxNzvO3pyvG39gC2r

Project team has approved the deliverables and the database has been taken into active use. Author has offered her help and assistance should any questions or problems arise.

7 Conclusion

The main aim of this thesis was to design and implement source data repository for the research project investigating institutional factors of knowledge-based economic development. Two key objectives were:

1. To identify and locate relevant data and data sources and set up a database solution to accommodate the source data;
2. To propose a preliminary subset of potential indicators for the research and build applications that enable further data pre-analysis and selection of final subsets of data for econometric models.

The key outcome of the thesis is a database implemented on MS Access platform, filled with research relevant data, which was identified as a result of a feature selection analysis conducted during this thesis. Database is supplemented with MS Access queries that facilitate the selection of final subsets of data for each econometric model. The results of the four sub-objectives as defined in section 1.3 are as follows:

1. Background of the research project, relevant economic theory and research into the works of other (institutional) authors investigating knowledge based economic development helped to shape the data requirements of the data repository (see chapter 3). Measurement of knowledge-based economy and development is a complex domain and typically involves measures of the quality and level of development of human resources, innovation system, (ICT) infrastructure, business environment, institutional efficiency and economic performance. An array of World Bank's sub-databases, along with other major international institutions, such as The Freedom House, Transparency International and Bertelsmann Foundation, that are at the forefront of gathering and producing reliable and internationally consistent country statistics, were selected to form the backbone of the database. Users of the database have access to more than 1500 indicators potentially relevant for the research project across more than 200 countries.

2. MS Access was identified as the most suitable solution for data storage based on the pre-set requirements (see chapter 4). Database, which resembles to a data lake in its nature, was designed and implemented incrementally in co-operation with the representative of the research team. Through various layers of the database selected data is rendered into a structured set of data suitable for analysis in data analysis software such as STATA, Eviews or R.

3. While final subsets of data for each econometric model shall be defined by the research team through iterative feature selection process, the author has carried out the first iteration of the feature selection process and proposes a pre-selection of 231 indicators, which have been selected considering the economic theory and background of the research as well as availability and relevance of the indicators in the database (see chapter 5.1 and Appendix 5).

4. Views, generated with SQL queries, provide comparative overview (with measures of descriptive statistics and availability) of the selected indicators and are useful for estimating the indicators' further suitability for the econometric models (see chapter 5.2.1 and 5.2.2). Modified version of the conformity analysis (Võhandu 1989; 2006), which is a data mining method allowing to identify groups of countries and indicators which are similar to each other in terms of data coverage, was selected as a complimentary tool to enable the data pre-analysis. Conformity analysis was implemented with MS Access 2016 SQL and MS Access 2016 VBA (see chapter 5.2.3).

In summary, author evaluates the main goals of the thesis to be met. The results of this thesis are readily usable, provide critical foundation to the whole research process and help to optimize further stages of the research. The database along with instructions (see Appendix 4) and key input material has been made available to the research team and to author's knowledge the database is in active use. Author is ready to assist the team whenever required.

8 Kokkuvõte

Antud töö peamiseks eesmärgiks oli disainida ja realiseerida andmebaas, mis koondab teadmuspõhise majandusarengu uurimisega seotud teadustöö jaoks vajalike algandmeid ning luua rakendused andmete eelanalüüsi hõlbustamiseks. Kaks põhieesmärki olid järgnevad:

1. Identifitseerida teadustööks olulised andmevaldkonnad ja peamised andmeallikad ning koondada andmed andmebaasi, mis on realiseeritud sobival platvormil;
2. Viia läbi andmete (indikaatorite ja riikide) eelvalik ning pakkuda välja rakendused, mis lihtsustavad andmete edasist eelanalüüsi ning võimaldavad genereerida ökonomeetriliseks modelleerimiseks sobivaid andmekogumeid paneelandmete formaadis.

Töö peamiseks väljundiks on MS Access platvormil realiseeritud andmebaas, mis koondab endas teadustöö jaoks huvipakkvaid andmeid. Andmebaasis on realiseeritud funktsionaalsused, mis lihtsustavad lõplike andmekogumike defineerimist ning eelanalüüsi läbiviimist. Järgnevalt on ära toodud töö tulemused alam-eesmärkide lõikes (peatüki 1.3 alusel):

1. Teadustöö taust, seotud majandusteooria ja teadmuspõhist majandusarengut uurivate (institutsionaalsete) autorite uuringute analüüs aitasid defineerida andmebaasile nõudeid (vt peatükk 3). Teadmuspõhise majandusarengu mõõtmine on kompleksne valdkond, hõlmates endas inimkapitali, innovatsioonisüsteemi, infrasturktuuri taset, ärikeskkonda, institutsionaalset efektiivsust ja üldist majadusarengut iseloomustavaid näitajaid. Andmebaas on ülesse ehitatud eelkõige Maailmapanga alamandmebaasides olevatele andmetele, kuid kasutatud on ka teiste rahvusvaheliselt tunnustatud ning usaldusväärset riikidepõhist statistikat avaldavate institutsioonide andmeid. Töö tulemusel on andmebaasi kasutajatel ligipääs rohkem kui 1500-le teadmuspõhise majadusarengu uurimiseks olulistele näidikutele rohkem kui 200 riigi lõikes.
2. MS Access hinnati kõige sobivamaks andmete koondamise platvormiks (vt peatükk 4). Andmebaas, mis oma olemuselt sarnaneb enim andmejärvele (ingl. k. *data lake*), realiseeriti inkrementaalselt koostöös teadusgrupi esindajaga.

Andmebaasi erinevad kontseptuaalsed kihid võimaldavad kogutud toorandmed transformeerida ökonomeetriliseks modelleerimiseks sobivasse paneelandmete formaati ning andmeid valitud andmeanalüüsi programmis koheselt analüüsida.

3. Kuigi lõplike andmekogumite defineerimine iga mudeli jaoks jääb projekti töörühma vastutada, siis töö autor viis läbi indikaatorite eelanalüüsi, hinnates kogutud andmete sobivust teadustöö eesmärkidest ja majandusteoreetilisest taustast lähtuvalt, ja pakub edasiseks analüüsiks 231 indikaatorit (vt peatükk 5.1 ja Lisa 5).

4. SQL päringutega realiseeritud vaated võimaldavad saada valitud indikaatoritest ülevaate (läbi kirjeldava statistika) ning hinnata nende sobivust edasiseks analüüsiks eri aspektidest lähtuvalt. Täiendavalt valiti eelanalüüsi lihtsustamiseks andmekaeve meetod konformsusanalüüs (Võhandu 1989; 2006), mis võimaldab tuvastada homogeensete indikaatorite ja riikide grupid andmete ajalisest katvusest lähtuvalt. Andmete eripärast tulenevalt tuli algset meetodit modifitseerida. Konformsusanalüüs realiseeriti MS Access 2016 SQL päringute ja MS Access 216 VBA-ga. (vt peatükk 5.2.3).

Kokkvõtteks hindab töö autor, et töö eemärgid saavutati. Töö tulemused on koheselt kasutatavad, koondatud andmed loovad teadustöö edasisteks etappideks olulise alusbaasi ja aitavad optimeerida teadustöö jägnevaid etappe. Andmebaas koos juhenditega (vt Lisa 4) on üle antud projekti töörühmale ja autorile teadaolevalt on need aktiivses kasutuses. Autor on vajadusel valmis töörühma andmebaasi kasutamisel igakülgselt abistama.

References

Acemolu, D., Johnson, S., & Robinson, J. (2005). Institutions as a fundamental cause of long run growth. In P. Edited by Aghion, & S. Durlauf, *Handbook of Economic Growth.* Elsevier B.V.

Arundel, A., Hansen, W., & Minna, K. (2008). *Indicators for the Knowledge-Based Economy: Summary Report.*

Asian Development Bank. (2007). *Moving Towards Knowledge-Based Economies: Asian Experiences.* Manila.

Asian Development Bank. (2014). *Innovative Asia: Advancing the Knowledge-Based Economy.* Manila, Philippines: Asian Development Bank.

Asia-Pacific Economic Cooperation. (2000). *Towards Knowledge-Based Economies in APEC.* APEC Economic Committee.

Barro, R. (1991). Economic Growth in a Cross-Section of Countries. *Quarterly Journal of Economics, 106*(2), 407-443.

Brooks, C. (2008). Introductory Econometrics for Finance. Cambridge University Press.

Chen, D., & Dahlman, C. (2005). *The Knowledge Economy, the KAM Methodology and World Bank Operations.* Washington DC: The World Bank.

Cohen, D., & Soto, M. (2001, September). Growth and Human Capital: Good Data, Good Results. *Technical Papers*(179).

Connolly, T., & Begg, C. (2002). *Database Systems: A practical Approach to Design, Implementation, and Management.* Harlow: Pearson Education Limited.

Date, J. (1997). *A guide to the SQL standard a user's guide to the standard database language SQL.* Addison-Wesley Professional.

Developer Information: Overview. (2017, May). Retrieved from The World Bank: https://datahelpdesk.worldbank.org/knowledgebase/articles/889386-developer-information-overview

Dull, T. (2017). Data Lake vs Data Warehouse: Key Differences Retrieved from: http://www.kdnuggets.com/2015/09/data-lake-vs-data-warehouse-key-differences.html

Eesaar, E. (2008). *Andmebaaside Projekteerimine.* Tallinn: TTÜ Kirjastus. Retrieved from www.eesti.ee

Elmasri, R. and Navathe, S. (2010). *Fundamentals of Database Systems*. Addison-Wesley Publishing Company.

Griffith, R., Redding, S., & Van Reenen, J. (2004). Mapping the Two Faces of R&D: Productivity Growth in a Panel of OECD industries. *Review of Economics and Statistics, 86*(4), 883-895.

Guellec, D., & Van Pottelsberghe de la Potterie, B. (2001). R&D and Productivity Growth: Panel Data Analysis of 16 OECD countries. *STI Working Paper, 2001/3*.

Gujarati, D. N. (2004). *Basic Econometrics.* The McGraw-Hill.

Günther, H. (2005). *Conference on Knowledge Economy: Challenges for Measurement.* Luxembourg.

Hazak, A. e. (forthcoming).

Kuusik, R., Lind, G. and Võhandu, L. (2004). Data Mining: Pattern Mining as a Clique Extracting Task. *ICEIS* (2) (pp. 519-522).

Lederman, D., Maloney, W. (2003). R&D and Development. *Policy Research Working Paper, 3024.*

Liiv, I., Kuusik, R., & Võhandu, L. (2007). Conformity analysis with structured query language. *Proceedings of the 6th International Conference on Artificial Intelligence, Knowledge Engineering and Databases*, (pp. 16-19).

Mo, P.H. (2001). Corruption and economic growth. *Journal of Comparative Economic*s, 29(1), pp.66-79.

Ehrlich, I., Lui, F.T. (1999). Bureaucratic corruption and endogenous economic growth. *Journal of Political Economy*, 107(S6), pp. 270-S293.

North, D. (1990). *Institutions, Institutional Change and Economic Performance.* New York: Cambridge University Press.

OECD. (1996). *The Knowledge Based Economy.* Paris: OECD.

OECD. (2001). *The New Economy: Beyond the Hype.* World Bank.

Oracle. (2017, May). *Oracle.* Retrieved from Database Data Warehousing Guide: https://docs.oracle.com/cd/B19306_01/server.102/b14223/ettover.htm

Oxford Dictionaries. (2017, 02 22). *Oxford Dictionaries*. Retrieved from https://en.oxforddictionaries.com/definition/institution

Paas, T. (1995). *Sissejuhatus ökonomeetriasse.* Tartu: Tartu Ülikooli kirjastus.

Piatetsky, G. (2004). *CRISP-DM, still the top methodology for analytics, data mining, or data science projects.* Retrieved from http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

Powell, W. W., & Snellmann, K. (2004). The Knowledge Economy. *Annual Review of Sociology*, 30:199-220.

Roman, J. (2016). CRISP-DM: The methodology to put some order into Data Science projects. Retrieved from https://data.sngular.team/en/art/40/crisp-dm-the-methodology-to-put-some-order-into-data-science-projects

Smith, K. (2000). *Innovation Indicators and the Knowledge Economy. Concepts, Results and Policy Challenges.* Oslo.

StataCorp LP. (2013). *STATA User's Guide Release 13.* Texas: Stata Press. Retrieved from http://www.stata.com/manuals13/u11.pdf

StatSoft, Inc. (2013). *Electronic Statistics Textbook*. Tulsa, OK: StatSoft. Retrieved from http://www.statsoft.com/textbook/.

Stewart, K. G. (2005). *Introduction to Applied Econometrics.* Thomson Learning.

The World Bank. (2008). *Measuring Knowledge in the World's Economies.* World Bank Institute.

Timmer, C. T. (2006). *How Countries Get Rash.* Washington: Centre for Global Development.

World Bank. (2016). *Vietnam 2035 : Toward Prosperity, Creativity, Equity, and Democracy.* World Bank.

World Bank Group. (2017). *Cambodia economic update : staying competitive through improving productivity.* World Bank Group.

World Bank Group. (2017). *Lao PDR Systematic Country Diagnostic.* Xieng Ngeun Village: World Bank Group.

Võhandu, L. (1989). Fast Methods in Exploratory Data Analysis. *Transactions of TTU,* 3-13.

Võhandu, L., Kuusik, R., Torim, A., Aab, E., & Lind, G. (2006). Some algorithms for data table (re) ordering using Monotone Systems. *Proceedings of the 5th WSEAS International Conference on Artificial Intelligence, Knowledge Engineering and Databases*, (p. 417-422).

Appendix 1 – Database base table descriptions

Table: indicator

Field Name	Description
ID	Row identifier.
is_downloaded	Enables the user to select indicators for data update from World Bank.
is_selected	Enables the user to select data (indicator/country/year) for output tables (master_cross_table, combined_indicator_values).
full_name	Indicator full name as defined by the source (all indicators from World Bank databank) or given names by the author (indicators from other data sources).
aggregates_to	Indicates the name of the parent indicator, if such exits. Usually these indicators are (sub)indexes and should be reviewed in conjunction with the parent index.
short_code	Short code given by the author for data manipulation and visualization purposes. Short codes are used in the data output tables to refer to indicators.
type	Scale type of the indicator (shows on what scale the indicator is measured). This information is useful for selecting analysis method. Following classification has been used: Binary - two categories (1/0) Nominal - unordered categories Ordinal - ordered categories, intervals between measurements are not meaningful (non-numeric) Numeric - numeric data on interval or ratio scale which is not classified under index, index100 and percent. Index - Indexes with values ranging from 0 to 10+ Index100 - Indexes with values ranging from 0 to 100+ Percent
category	Indicates the category where the indicator has been classified. Classification is tentative. Classes represent logical groupings of the knowledge-based economy indicators. Following classification has been used: STRUCT - Structural DEMO - Demographic HEALTH - Health EDU - Education and quality of human resources ECONPER - Economic performance INNOSYS - Innovation System ICTINFRA - ICT infrastructure BUSENV - Business environment INST - Institutional regime and efficiency
data_source	Institution where the data has been obtained.
database_name	Subdatabase (database, project or similar) of the data source, if available.
wb_code	World Bank official indicator codes, available only for indicators from World Bank.
wb_topic	Topic under which World Bank has classified the indicator, available only for World Bank indicators.
definition	Definitions provided by the source (all World Bank indicators) or by the author (all other sources).
aggregation_method	Method by which the aggregation has been obtained provided by the source (all World Bank indicators) or by the author (all other sources), where possible.

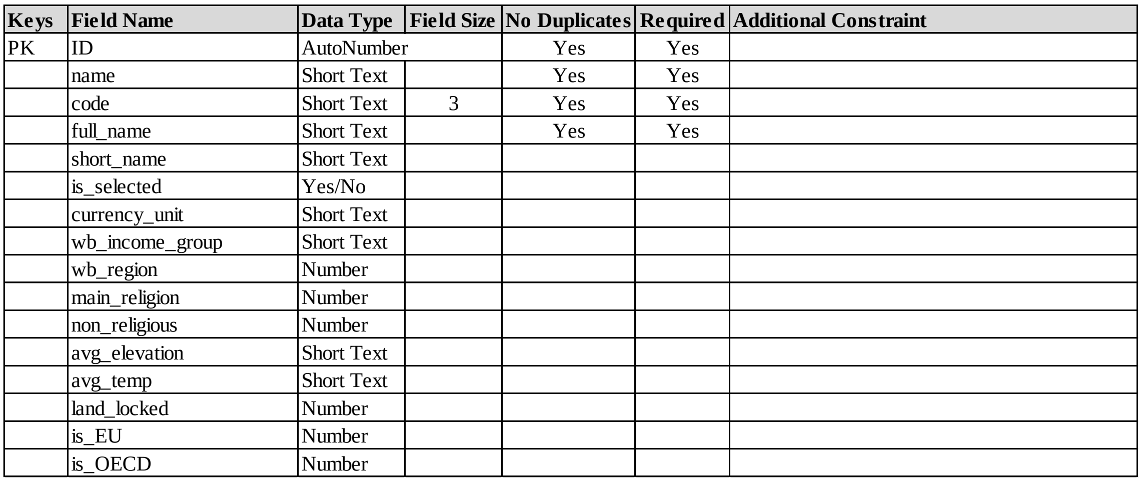
Field Name	Description
source_description	Describes the institution/data source of the indicator (in case of World Bank indicators it refers to the initial data source where World Bank has obtained the data).
stat_concept_methodology	Describes the statistical concept and methodology used to compute the indicator values, if available, provided by the source (all World Bank indicators) or by the author (all other sources).
limitations_exceptions	Describes the limitations and exceptions of the indicators (including the shortcomings in its methodology) the users of the data should be aware of, if available, provided by the source (all World Bank indicators) or by the author (all other sources).

Keys	Field Name	Data Type	Field Size	No duplicates	Required	Additional constraint
PK	ID	AutoNumber		Yes	Yes	
	is_downloaded	Yes/No				
	is_selected	Yes/No				is_selected=Yes ONLY IF is_downloaded=Yes
	full_name	Short Text		Yes	Yes	
	aggregates_to	Short Text				
	short_code	Short Text	8	Yes		NOT NULL when is_selected=Yes
	type	Short Text				
	category	Short Text				
	data_source	Short Text				
	database_name	Short Text				
	wb_code	Short Text				
	wb_topic	Short Text				
	definition	Long Text				
	aggregation_method	Short Text				
	source_description	Short Text				
	stat_concept_methodology	Long Text				
	limitations_exceptions	Long Text				

Table: country

Field Name	Description
ID	Row identifier.
is_downloaded	Enables the user to select indicators for data update from World Bank.
is_selected	Enables the user to select data (indicator/country/year) for output tables (master_cross_table, combined_indicator_values).
full_name	Indicator full name as defined by the source (all indicators from World Bank databank) or given names by the author (indicators from other data sources).
aggregates_to	Indicates the name of the parent indicator, if such exits. Usually these indicators are (sub)indexes and should be reviewed in cojunction with the parent index.
short_code	Short code given by the author for data manipulation and visualization purposes. Short codes are used in the data output tables to refer to indicators.
type	Scale type of the indicator (shows on what scale the indicator is measured). This information is useful for selecting analysis method. Following classification has been used: Binary - two categories (1/0) Nominal - unordered categories Ordinal - ordered categories, intervals between measurements are not meaningful (non-numeric) Numeric - numeric data on interval or ratio scale which is not classified under index, index100 and percent. Index - Indexes with values ranging from 0 to 10+ Index100 - Indexes with values ranging from 0 to 100+ Percent

Field Name	Description
category	Indicates the category where the indicator has been classified. Classification is tentative. Classes represent logical groupings of the knowledge-based economy indicators. Following classification has been used: STRUCT - Structural DEMO - Demographic HEALTH - Health EDU - Education and quality of human resources ECONPER - Economic performance INNOSYS - Innovation System ICTINFRA - ICT infrastructure BUSENV - Business environment INST - Institutional regime and efficiency
data_source	Institution where the data has been obtained.
database_name	Subdatabase (database, project or similar) of the data source, if available.
wb_code	World Bank official indicator codes, available only for indicators from World Bank.
wb_topic	Topic under which World Bank has classified the indicator, available only for World Bank indicators.
definition	Definitions provided by the source (all World Bank indicators) or by the author (all other sources).
aggregation_method	Method by which the aggregation has been obtained provided by the source (all World Bank indicators) or by the author (all other sources), where possible.
source_description	Describes the institution/data source of the indicator (in case of World Bank indicators it refers to the initial data source where World Bank has obtained the data).
stat_concept_methodology	Describes the statistical concept and methodology used to compute the indicator values, if available, provided by the source (all World Bank indicators) or by the author (all other sources).
limitations_exceptions	Describes the limitations and exceptions of the indicators (including the shortcomings in its methodology) the users of the data should be aware of, if available, provided by the source (all World Bank indicators) or by the author (all other sources).



Keys	Field Name	Data Type	Field Size	No Duplicates	Required	Additional Constraint
PK	ID	AutoNumber		Yes	Yes	
	name	Short Text		Yes	Yes	
	code	Short Text	3	Yes	Yes	
	full_name	Short Text		Yes	Yes	
	short_name	Short Text				
	is_selected	Yes/No				
	currency_unit	Short Text				
	wb_income_group	Short Text				
	wb_region	Number				
	main_religion	Number				
	non_religious	Number				
	avg_elevation	Short Text				
	avg_temp	Short Text				
	land_locked	Number				
	is_EU	Number				
	is_OECD	Number				

Table: year

Field Name	Description
ID	Row identifier
year	List of year values across which data can be available (data downloaded from World Bank is by default starting from 1960 till latest available).
is_selected	Enables the user to choose years of interest into the final output tables.

Keys	Field Name	Data Type	Field Size	No Duplicates	Required	Additional Constraint
PK	ID	Autonumber		Yes	Yes	
	date	Short Text		Yes	Yes	
	is_selected	Short Text				

Table: combined_indicator_values

Field Name	Description
ID	Row identifier
indicator	Refers to the column ‘full_name’ in table Indicator.
country	Refers to the column ‘name’ in table Country.
date	Refers to the column ‘date’ in table Year.

Keys	Field Name	Data Type	Field Size	No Duplicates	Required	Additional Constraint
PK	ID	Autonumber		Yes	Yes	
FK	indicator	Short Text				
FK	country	Short Text				
FK	date	Short Text				

Appendix 2 – Modules

Module: Pull Data

```
*** Extracting data from World Bank API

Option Compare Database
Function PullData()

Dim dbs As DAO.Database
Dim indicatorList, countrieInfoList As DAO.Recordset

Set dbs = CurrentDb
Set indicatorList = dbs.OpenRecordset("indicator", dbOpenTable)

Dim seriesCode As String
Dim addIndicator As Boolean

Dim firstQuery As Boolean
Dim sourceDatabase As String
firstQuery = True

If doesTableExist("data") Then
    dbs.Execute ("DROP TABLE data")
End If

If doesTableExist("wb_indicator_values") Then
    dbs.Execute ("DROP TABLE wb_indicator_values")
End If

Do Until indicatorList.EOF = True


    addIndicator = indicatorList!is_downloaded
    If IsNull(indicatorList!data_source) Then
        sourceDatabase = ""
    Else
        sourceDatabase = indicatorList!data_source
    End If

    If addIndicator And sourceDatabase = "World Bank" Then
        seriesCode = indicatorList!wb_code
        If firstQuery Then
            On Error GoTo handleError
            Application.ImportXML
DataSource:="http://api.worldbank.org/countries/all/indicators/" + seriesCode
+ "?per_page=20000", ImportOptions:=acStructureAndData
            On Error GoTo 0
            firstQuery = False
        Else
            On Error GoTo handleError
            Application.ImportXML
DataSource:="http://api.worldbank.org/countries/all/indicators/" + seriesCode
+ "?per_page=20000", ImportOptions:=acAppendData
            On Error GoTo 0
        End If
Continue:
    End If
    indicatorList.MoveNext
Loop

Dim filteredSet As Recordset
Dim strSQL As String

strSQL = "SELECT data.indicator,data.country,data.date,data.value INTO
wb_indicator_values FROM data INNER JOIN country ON
data.country=country.name;"
dbs.Execute (strSQL)

If doesTableExist("data") Then
    dbs.Execute ("DROP TABLE data")
End If

dbs.Close
Exit Function

handleError:
  Dim result As Integer
  result = MsgBox(Err.Description & "The indicator was: " & seriesCode, _
    vbExclamation + vbOKCancel, _
    "Error: " & CStr(Err.Number))
  If result = 2 Then
    Exit Function
  End If
Resume Continue
End Function

Public Function doesTableExist(strTableName As String) As Boolean
    Dim db As DAO.Database
    Dim td As DAO.TableDef
    Set db = CurrentDb
    On Error Resume Next
    Set td = db.TableDefs(strTableName)
    doesTableExist = (Err.Number = 0)
    Err.Clear
End Function
```

Module: Select and Combine

```
*** Generating combined dataset

Option Compare Database
Option Explicit
Function SelectCombineData()

Dim dbs As DAO.Database
Set dbs = CurrentDb

If doesTableExist("wb_fk_indicator_values") Then
    dbs.Execute ("DROP TABLE wb_fk_indicator_values")
End If

If doesTableExist("wb_indicator_values_f") Then
    dbs.Execute ("DROP TABLE wb_indicator_values_f")
End If

If doesTableExist("other_sources_indicator_values_f") Then
```

```
    dbs.Execute ("DROP TABLE other_sources_indicator_values_f")
End If

If doesTableExist("combined_indicator_values") Then
    dbs.Execute ("DELETE combined_indicator_values.* FROM
combined_indicator_values")
End If

CurrentDb.Execute "1_decrease_data_volume"
CurrentDb.Execute "2_change_datatype_to_double"
CurrentDb.Execute "3_make_os_final_dataset"
CurrentDb.Execute "3_make_wb_final_dataset"
CurrentDb.Execute "4_make_combined_dataset"

If doesTableExist("wb_fk_indicator_values") Then
    dbs.Execute ("DROP TABLE wb_fk_indicator_values")
End If

If doesTableExist("wb_indicator_values_f") Then
    dbs.Execute ("DROP TABLE wb_indicator_values_f")
End If

If doesTableExist("other_sources_indicator_values_f") Then
    dbs.Execute ("DROP TABLE other_sources_indicator_values_f")
End If

‘rule validation messages
Dim codeList As DAO.Recordset
Dim downloadList As DAO.Recordset
Dim countMissingDl As Integer
Dim countMissingCode As Integer

Set downloadList = dbs.OpenRecordset("SELECT count(full_name) As
count_missing FROM( SELECT
DISTINCT([indicator].full_name),combined_indicator_values.[indicator]FROM
[indicator] LEFT JOIN combined_indicator_values ON
[indicator].full_name=combined_indicator_values.[indicator]WHERE
[indicator].is_selected=TRUE)WHERE [indicator] IS NULL")
countMissingDl = downloadList!count_missing
If countMissingDl > 0 Then
   MsgBox ("Source data has not been downloaded for " & countMissingDl & "
selected indicators!" & vbNewLine & "Output tables may not return desired
results." & vbNewLine & "Please de-select indicators which have not been
downloaded or download indicators first! ")
End If
downloadList.Close

Set codeList = dbs.OpenRecordset("SELECT count(full_name) As
short_code_missing FROM [indicator] WHERE is_selected = True And
IsNull(short_code)")
countMissingCode = codeList!short_code_missing
If countMissingCode > 0 Then
   MsgBox ("Short code is missing for " & countMissingCode & " selected
indicators!" & vbNewLine & "Output tables may not return desired results." &
vbNewLine & "Please assign short codes in table Indicator! ")
End If
codeList.Close

dbs.Close
```

```
End Function

Public Function doesTableExist(strTableName As String) As Boolean
    Dim db As DAO.Database
    Dim td As DAO.TableDef
    Set db = CurrentDb
    On Error Resume Next
    Set td = db.TableDefs(strTableName)
    doesTableExist = (Err.Number = 0)
    Err.Clear
End Function
```

Module: Update Master Table

```
***Obtaining ranking of countries

Option Compare Database
Option Explicit

Function UpdateMasterCrossTable()

Dim dbs As DAO.Database
Set dbs = CurrentDb

If doesTableExist("master_cross_table") Then
    dbs.Execute ("DROP TABLE master_cross_table")
End If

Dim strSQL As String
strSQL = "SELECT query_master_cross_table.* INTO master_cross_table FROM
query_master_cross_table;"
dbs.Execute (strSQL)

dbs.Close
End Function

Public Function doesTableExist(strTableName As String) As Boolean
    Dim db As DAO.Database
    Dim td As DAO.TableDef
    Set db = CurrentDb
    On Error Resume Next
    Set td = db.TableDefs(strTableName)
    doesTableExist = (Err.Number = 0)
    Err.Clear
End Function
```

Module: Conformity Country

```
***Obtaining ranking of countries

Option Compare Database
Option Explicit

Function Minus_IterationsCountry()
Dim dbs As DAO.Database
Set dbs = CurrentDb

Dim number_of_states As Long
Dim i As Long

If doesTableExist("c_combined_indicator_values") Then
```

```
    dbs.Execute ("DROP TABLE c_combined_indicator_values")
End If

If doesTableExist("conformity_country_rank") Then
    dbs.Execute ("DELETE conformity_country_rank.* FROM
conformity_country_rank")
End If

CurrentDb.Execute "c_copy_combined"

number_of_states = DCount("country", "c_combined_indicator_values")
For i = 1 To number_of_states
    CurrentDb.Execute "c_save_iteration_country"
    CurrentDb.Execute "c_delete_iteration_country_records"
Next i
CurrentDb.Execute ("DROP TABLE c_combined_indicator_values")
dbs.Close
End Function

Public Function doesTableExist(strTableName As String) As Boolean
    Dim db As DAO.Database
    Dim td As DAO.TableDef
    Set db = CurrentDb
    On Error Resume Next
    Set td = db.TableDefs(strTableName)
    doesTableExist = (Err.Number = 0)
    Err.Clear
End Function
```

Module: Conformity Indicator

***Obtaining ranking of indicators

```
Option Compare Database
Option Explicit

Function Minus_IterationsIndicator()

Dim dbs As DAO.Database
Set dbs = CurrentDb

Dim number_of_indicators As Long
Dim i As Long

If doesTableExist("c_combined_indicator_values") Then
    dbs.Execute ("DROP TABLE c_combined_indicator_values")
End If

If doesTableExist("conformity_indicator_rank") Then
    dbs.Execute ("DELETE conformity_indicator_rank.* FROM
conformity_indicator_rank")
End If

CurrentDb.Execute "c_copy_combined"

number_of_indicators = DCount("short_code", "c_combined_indicator_values")
For i = 1 To number_of_indicators
    CurrentDb.Execute "c_save_iteration_indicator"
    CurrentDb.Execute "c_delete_iteration_indicator_records"
Next i
```

```
CurrentDb.Execute ("DROP TABLE c_combined_indicator_values")
dbs.Close
End Function

Public Function doesTableExist(strTableName As String) As Boolean
    Dim db As DAO.Database
    Dim td As DAO.TableDef
    Set db = CurrentDb
    On Error Resume Next
    Set td = db.TableDefs(strTableName)
    doesTableExist = (Err.Number = 0)
    Err.Clear
End Function
```

Module: Conformity Combine

```
***Obtaining ranking of indicators

Option Compare Database
Option Explicit
Function Minus_IterationsCombine()
Dim dbs As DAO.Database
Set dbs = CurrentDb
If doesTableExist("conformity_cross_table") Then
    dbs.Execute ("DROP TABLE conformity_cross_table")
End If
Dim strSQL As String
strSQL = "SELECT query_conformity_cross_table.* INTO conformity_cross_table
FROM query_conformity_cross_table;"
dbs.Execute (strSQL)
dbs.Close
End Function

Public Function doesTableExist(strTableName As String) As Boolean
    Dim db As DAO.Database
    Dim td As DAO.TableDef
    Set db = CurrentDb
    On Error Resume Next
    Set td = db.TableDefs(strTableName)
    doesTableExist = (Err.Number = 0)
    Err.Clear
End Function
```

Appendix 3 – Queries

Queries generating views for analysis layer

View1_selected_indicator_stats

```
SELECT [indicator].category, [indicator].data_source,
[indicator].database_name AS subdatabase,
combined_indicator_values.[indicator], [indicator].short_code,
[indicator].type AS scale_type, COUNT([value]) AS nr_of_obs, min([date]) AS
first_year, max([date]) AS last_year, min([value]) AS min_value, max([value])
AS max_value, avg([value]) AS av, STDEV([value]) AS stdev
FROM combined_indicator_values LEFT JOIN [indicator] ON
combined_indicator_values.[indicator]=[indicator].full_name
GROUP BY [indicator].category, [indicator].data_source,
[indicator].database_name, combined_indicator_values.[indicator],
[indicator].short_code, [indicator].type
ORDER BY [indicator].category, [indicator].database_name;
```

View2_indicator_period_coverage

```
TRANSFORM First(a.[value]) AS FirstOfvalue
SELECT [indicator].category AS indicator_category, a.indicator AS
indicator_name, [indicator].short_code, a.country, a.code AS country_code
FROM (SELECT combined_indicator_values.*, country.code FROM
combined_indicator_values LEFT JOIN country ON
country.name=combined_indicator_values.country)  AS a LEFT JOIN [indicator]
ON a.[indicator]=[indicator].full_name
GROUP BY [indicator].category, a.[indicator], [indicator].short_code,
a.country, a.code
PIVOT a.[date];
```

Queries used for pulling and updating data selection and updating master output table

1_decrease_data_volume

```
SELECT [indicator].ID, country.ID, wb_indicator_values.[date],
wb_indicator_values.[value] INTO wb_fk_indicator_values
FROM ([indicator] INNER JOIN (country INNER JOIN wb_indicator_values ON
country.name = wb_indicator_values.country) ON [indicator].full_name =
wb_indicator_values.[indicator]) INNER JOIN [year] ON
wb_indicator_values.[date] = [year].[date]
WHERE ((([year].is_selected)=Yes) AND ((country.is_selected)=Yes) AND
(([indicator].is_selected)=Yes));
```

2_change_datatype_to_double

```
ALTER TABLE wb_fk_indicator_values ALTER COLUMN [value] DOUBLE;
```

3_make_os_final_dataset

```
SELECT other_sources_indicator_values.[indicator],
other_sources_indicator_values.country,
other_sources_indicator_values.[date], other_sources_indicator_values.[value]
INTO other_sources_indicator_values_f
FROM ((other_sources_indicator_values INNER JOIN [year] ON
other_sources_indicator_values.[date] = [year].[date]) INNER JOIN country ON
other_sources_indicator_values.country = country.name) INNER JOIN [indicator]
ON other_sources_indicator_values.[indicator] = [indicator].full_name
WHERE ((([year].is_selected)=Yes) AND ((country.is_selected)=Yes) AND
(([indicator].is_selected)=Yes));
```

3_make_wb_final_dataset

```
SELECT [indicator].full_name AS [indicator], country.name AS country,
wb_fk_indicator_values.[date] AS [date], wb_fk_indicator_values.[value] AS
[value] INTO wb_indicator_values_f
FROM (wb_fk_indicator_values INNER JOIN [indicator] ON
wb_fk_indicator_values.indicator_ID = [indicator].ID) INNER JOIN country ON
wb_fk_indicator_values.country_ID = country.ID;
```

4_make_combined_dataset

```
INSERT INTO combined_indicator_values ( [indicator], country, [date], [value]
)
SELECT a.[indicator], a.country, a.[date], a.[value]
FROM (SELECT * FROM wb_indicator_values_f UNION ALL SELECT * FROM
other_sources_indicator_values_f)  AS A LEFT JOIN [indicator] ON
[indicator].full_name=a.[indicator];
```

query_master_cross_table

```
TRANSFORM First(a.[value]) AS FirstOfvalue
SELECT a.country, a.code, a.[date] AS [year], a.main_religion AS religion,
a.non_religious AS non_rel, a.land_locked AS is_locked, a.avg_elevation AS
avg_elev, a.avg_temp
FROM (SELECT combined_indicator_values.*, country.code,
country.main_religion, country.non_religious, country.land_locked,
country.avg_elevation, country.avg_temp FROM combined_indicator_values LEFT
JOIN country ON country.name=combined_indicator_values.country WHERE
country.is_selected = True)  AS a LEFT JOIN [indicator] ON
a.[indicator]=[indicator].full_name
GROUP BY a.country, a.code, a.[date], a.main_religion, a.non_religious,
a.land_locked, a.avg_elevation, a.avg_temp
PIVOT [indicator].short_code;
```

Queries used in conformity analysis

c_copy_combined

```
SELECT [indicator], country, [date], [value], [indicator].short_code,
[indicator].category INTO c_combined_indicator_values
FROM combined_indicator_values LEFT JOIN [indicator] ON
combined_indicator_values.[indicator]=[indicator].full_name;
```

c_country_frequency

```
SELECT short_code, [date] AS [year], count(value) AS country_frequency
FROM c_combined_indicator_values
GROUP BY short_code, [date];
```

c_country_sum

```
SELECT country, sum(country_frequency) AS country_sum
FROM c_data_table_freq
WHERE value<>NULL
GROUP BY country
ORDER BY sum(country_frequency);
```

c_country_sum_TOP1

```
SELECT TOP 1 c_country_sum.country, c_country_sum.country_sum
FROM c_country_sum;
```

c_data_table_freq

```
SELECT c_combined_indicator_values.country,
c_combined_indicator_values.[date] AS [year], c_country_frequency.short_code,
c_combined_indicator_values.[value], c_country_frequency.country_frequency
FROM c_combined_indicator_values LEFT JOIN c_country_frequency ON
(c_combined_indicator_values.short_code = c_country_frequency.short_code) AND
(c_combined_indicator_values.[date] = c_country_frequency.[year]);
```

c_data_table_freq_ind

```
SELECT c_combined_indicator_values.country,
c_combined_indicator_values.[date] AS [year],
c_combined_indicator_values.short_code, c_combined_indicator_values.[value],
c_indicator_frequency.indicator_frequency
FROM c_combined_indicator_values LEFT JOIN c_indicator_frequency ON
(c_combined_indicator_values.[date]=c_indicator_frequency.[year]) AND
(c_combined_indicator_values.country=c_indicator_frequency.country);
```

c_indicator_frequency

```
SELECT country, date AS [year], count([value]) AS indicator_frequency
FROM c_combined_indicator_values
GROUP BY country, [date];
```

c_indicator_sum

```
SELECT short_code, sum(indicator_frequency) AS indicator_sum
FROM c_data_table_freq_ind
WHERE value<>NULL
GROUP BY short_code
ORDER BY sum(indicator_frequency);
```

c_indicator_sum_TOP1

```
SELECT TOP 1 c_indicator_sum.short_code, c_indicator_sum.indicator_sum
FROM c_indicator_sum;
```

c_value_frequency

```
SELECT short_code, country, count([value]) AS value_freq
FROM combined_indicator_values LEFT JOIN [indicator] ON
[indicator].full_name=combined_indicator_values.[indicator]
GROUP BY short_code, country;
```

c_delete_iteration_country_records

```
DELETE c_combined_indicator_values.*
FROM c_combined_indicator_values
WHERE
(((c_combined_indicator_values.country)=DLookUp("country","c_country_sum_TOP1
")));
```

c_delete_iteration_indicator_records

```
DELETE c_combined_indicator_values.*
FROM c_combined_indicator_values
WHERE
(((c_combined_indicator_values.short_code)=DLookUp("short_code","c_indicator_
sum_TOP1")));
```

c_save_iteration_country

```
INSERT INTO conformity_country_rank ( country, rank, Score )
SELECT b.country, a.rank, b.country_sum
FROM (SELECT count(c_country_sum.country) AS rank FROM c_country_sum)  AS a,
(SELECT TOP 1 * FROM c_country_sum)  AS b;
```

c_save_iteration_indicator

```
INSERT INTO conformity_indicator_rank ( rank, [indicator], score )
SELECT a.rank AS rank, b.short_code AS [indicator], b.indicator_sum AS score
FROM (SELECT count(c_indicator_sum.short_code) AS rank FROM c_indicator_sum)
AS a, (SELECT TOP 1 * FROM c_indicator_sum)  AS b;
```

query_conformity_cross_table

```
TRANSFORM First(b.value_freq) AS FirstOfvalue
SELECT country_rankf
FROM (SELECT b.*, conformity_indicator_rank.[indicator],
conformity_indicator_rank.rank AS indicator_rank,
Format(conformity_indicator_rank.rank,"000") & "_" &
conformity_indicator_rank.[indicator] AS indicator_rankf FROM
conformity_indicator_rank INNER JOIN (SELECT a.*,
conformity_country_rank.country, conformity_country_rank.rank AS
country_rank, Format(conformity_country_rank.rank,"000") & "_" &
conformity_country_rank.country AS country_rankf FROM conformity_country_rank
INNER JOIN (SELECT c.*, c_value_frequency.value_freq FROM (SELECT
combined_indicator_values.*, [indicator].short_code, [indicator].category
FROM combined_indicator_values LEFT JOIN [indicator] ON
[indicator].full_name=combined_indicator_values.[indicator])  AS c LEFT JOIN
c_value_frequency ON (c.short_code=c_value_frequency.short_code) AND
(c.country=c_value_frequency.country))  AS a ON
a.country=conformity_country_rank.country)  AS b ON
b.short_code=conformity_indicator_rank.[indicator] ORDER BY b.country_rank
DESC , conformity_indicator_rank.rank DESC)  AS c
GROUP BY country_rankf
PIVOT indicator_rankf;
```

Queries for restoring indicator and country pre-selection done by the author

Restore_country_pre_selection

```
UPDATE country INNER JOIN pre_selected_countries ON
pre_selected_countries.name=country.name SET country.is_selected = TRUE
WHERE pre_selected_countries.name=country.name;
```

Restore_indicator_pre_selection

```
UPDATE [indicator] INNER JOIN pre_selected_indicators ON
pre_selected_indicators.full_name=[indicator].full_name SET
[indicator].is_selected = TRUE
WHERE pre_selected_indicators.full_name=[indicator].full_name;
```

Queries for clearing indicator and country selection done by user

clear_country_selection

```
UPDATE country SET is_selected = FALSE;
```

clear_indicator_selection

```
UPDATE [indicator] SET is_selected = FALSE;
```

Appendix 4 – Instructions for the database

This database is designed to assist researchers working towards the "Institutions for Knowledge Intensive Development: Economic and Regulatory Aspects in South-East Asian Transition Economies" research project. Database contains data on diverse set of development indicators across all world countries, including Cambodia, Laos and Vietnam – countries at the focus of the research project.

Before operating the database, it is recommended to save a separate copy of the database.

Elements of the database:

1. Tables represent the permanent data tables.

    - *country* holds a complete list of world countries and economic territories (217) as defined by the World Bank and key structural time invariant data about the countries.

    - *indicator* holds a list of development indicators (1628) (based on World Development Indicators subdatabase), sourced mostly from World Bank.

    - *master_cross_table* represents the output table where data from selected countries and indicators has been combined into a format suitable (panel data) for analysis with data analysis programs. The contents of this table are renewed with macro 3.

    - *wb_indicator_values* holds indicator value data in retrieved from World Bank. Data in this table is updated when macro number 1 has successfully completed.

    - *other_sources_indicator_values* holds indicator value data from all other data sources. This table has been created and filled with data imported from Excel.

    - *combined_indicator_values* combines data from tables *wb_indicator_values* and *other_sources_indicator_values* into unanimous format based on the selections made in tables *country*, *indicator* and *year*. This table is updated with macro 2. This table forms the basis for *master_cross_table*.

    - *conformity_country_rank* ranks countries based on their 'conformity' score (1- best). This table is updated with macro number 4.

    - *conformity_indicator_rank* ranks indicators based on their 'conformity' score (1- best). This table is updated with macro number 5.

    - *conformity_cross_table* is data matrix, which is ordered based on the country and indicator rankings, useful for visually detecting homogeneous groups of indicators and countries (in terms of observation frequency). This table is updated with macro number 6.

- *pre-selected_countries* holds filtered list of countries, excluding small island nations and disputed territories, and forms basis for the *pre_selected_master_cross_table.*. Users are encouraged to form their own subsets based on this pre-selection. This table is of informative nature to showcase the pre-selection and it is not updateable (users can however implement this pre-selection in their iterations with query *restore_country_pre_selection*) .

- *pre-selected_indicators* holds the list of pre-selected indicators, which have been identified as relevant through first iteration of feature selection process, and forms basis for the *pre_selected_master_cross_table*. Users are encouraged to form their own subsets based on this pre-selection. This table is of informative nature to showcase the pre-selection and it is not updateable (users can however implement this pre-selection in their iterations with query *restore_indicator_pre_selection*).

- *pre_selected_master_cross_table* is the key output table in panel format formed based on the pre-selection of indicators and countries as saved in tables *pre-selected_indicators and pre-selected_countries*. This table is of informative nature to showcase the pre-selection and it is not updateable.

2. Queries, which are either part of macros or designed to provide overview of the selected data (views).

3. Macros to operate the database.

   (1) initiate the data pull from World Bank API, as a result contents of table *wb_indicator_values* are overwritten;

   (2) update the contents of *combined_indicator_values* based on the selections made in tables *year*, *country* and *indicator*;

   (3) overwrite the final output table *master_cross_table* (based on data contained in combined_indicator_values);

   (4-6) perform conformity analysis on the selected data. This analysis is performed on the data contained in table *combined_indicator_values*.

4. Modules, small programs written in VBA, initiated by the macros. Source code of these modules can be examined and modified in the design view.

Instructions to key operations:

Central functionality of the database is the ability to form sets of panel data across array of indicators, countries and time periods. While data was pulled to the database from World Bank API the success of the subsequent pulls is dependent on the stability of the indicator name definitions, which provide basis to the pull.

1. Updating source data

On delivery database holds indicator value data regarding 231 pre-selected and downloaded indicators. This list should be sufficient for the research. If, user wants to get some additional data or update data values (as new datapoint have become available), it can be done with macro 1. If this is the case, run macro number 1 to pull (or update) and indicator values from World Bank (per selection performed in step 2). **This can take more than 1 hour** depending on the connection speed and the number of indicators selected (pull for 231 indicators took around 1 hour). Message box will be displayed if the pull has been successful. It can happen that some indicators are no longer available. If this is the case, an error message will be displayed indicating the indicator (code) that could not be pulled. Pull is initiated based on the indicators marked as 'is_downloaded' in table *indicator*.

Data which source is not World Bank cannot be updated automatically.

2. Updating the final output table master_cross_table

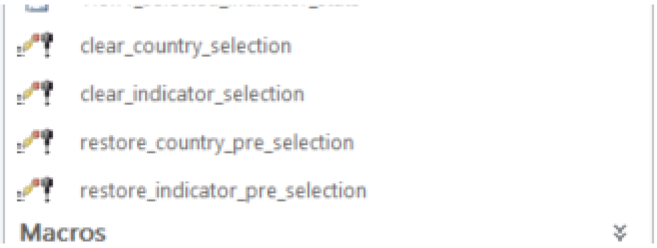
1. Select desired countries from table *country* by ticking the column "is_selected" and desired time-period in table *year* by ticking the column "is_selected". Save the table and **close the table!** (Advisable not to exceed the pre-selection of countries of 152 and period of 30 years, Access may not be able to save it as a table due to column limitations).

2. Select desired indicators from table *indicator* by ticking the column "is_selected". The key logic here is that indicators, which have been downloaded are only selectable, i.e they must be 'is_downloaded'=Yes and the pull macro 1 needs to be initiated. If this rule is violated the output tables simply wont be able to display this data. Each new indicator downloaded into the database and included into the selection 'is_selected'=Yes needs to be given unique 'short_code' because 'short_code' is used to display indicators in panel format. User is prompted if this data is missing. Save the table and **close the table!** (Advisable not to exceed the pre/selection of indicators of 231)

3. After selections are done run macro 2 which combines data from tables *other_sources_indicator_values* and *wb_indicator_value*s into *combined_indicator_values*. Initiate macro 2 every time you wish to overwrite the output tables based on new *country, year*, *indicator* selection. You also need to run macro 3 (see next).

4. Run macro 3 to update the *master_cross_table* and overwrite the table based on the updated *combined_indicator_valu*es table. **This can take up to 10 min depending** on the number of indicators selected. Message box will be displayed if the update has been successful. *master_cross_table* data has been updated and conforms to the country and indicator selection. If too many indicators have been selected, Access will not be able to save it as a table. Either decrease indicators in selection or use the query to produce the *view* and transfer the data from the *view* to data analysis program.

NB! Indicator and country selection can be cleared with queries *clear_country_selection* and *clear_indicator_selection* (see below), all 'is_selected' columns will be set to 'No'.



clear_country_selection

clear_indicator_selection

restore_country_pre_selection

restore_indicator_pre_selection

Macros

Similarly, user can restore the pre-selection proposed by the author with queries *restore_country_pre_selection* and *restore_indicator_pre_selection*.

3. Data pre-analysis

There are several functionalities designed to help assess the data selected and through iterations define finals sets of indicators and countries which are imported to data analysis program.

1. Query=> *view1_selected_indicator_*stats provides descriptive stats regarding the selected indicators.

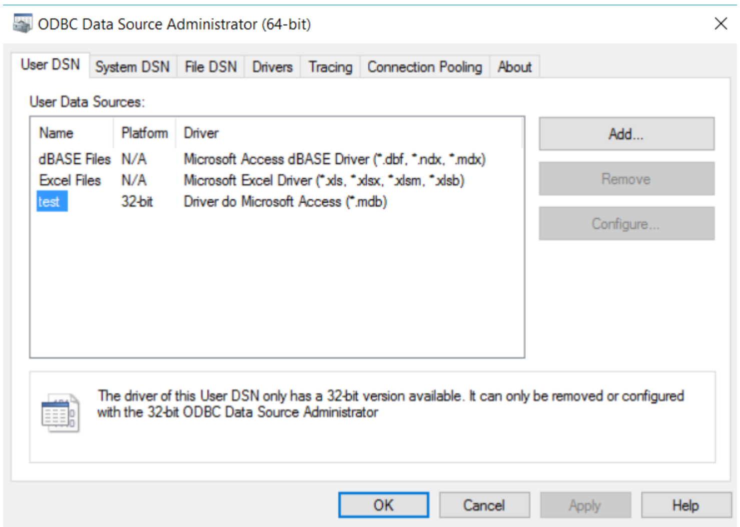
2. Query=> *view2_indicator_period_coverge* provides temporal overview of the selected indicators across selected countries.

3. Conformity analysis – the output of the analysis is generated based on the years, countries and indicators which have been ticked off 'is_selected' in relevant source tables. First select desired years, countries and indicator (not advisable to operate with more than 15x15 matrices) save tables, **close the tables** and run macro nr 2 to update data selection and then macro nr 4, 5 and 6 to overwrite the analysis results into table *conformity_cross_table*. The conformity analysis update can take up long time depending on the number of variables selected (for instance, analysis with 5 countries, 7 indicators and 30 years takes up around 5 minutes). The analysis steps have been divided into individual steps because in case of big data selections database may need to be compacted and repaired (File=>Compact and Repair) in between these steps in order to be successful. This is caused by the capacity limitations of 2 GB MS Access.

Importing data to data analysis software

There are two main options for data import to data analysis software. First through Excel export, which might be the easiest. Second option is to set up permanent connection over ODBC. The main steps in setting up the ODBC connection are as follows:

1. Ensure you have set up data source name (DSN) in the OBDC Data Source Administrator (Control Panel=>System and Security=>Administrative Tools=>Data Sources (ODBC). Once the 'Source Administrator' window pops up, click 'Add' and select appropriate driver from the list (MS Access). Define data source name and select the database you wish to connect to.



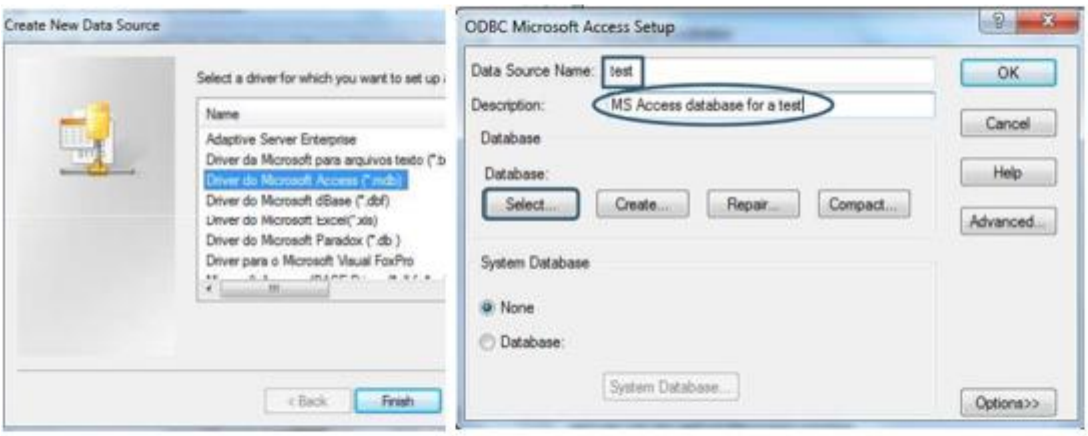
ODBC Data Source Administrator (64-bit)

User DSN | System DSN | File DSN | Drivers | Tracing | Connection Pooling | About

User Data Sources:

Name	Platform	Driver
dBASE Files	N/A	Microsoft Access dBASE Driver (*.dbf, *.ndx, *.mdx)
Excel Files	N/A	Microsoft Excel Driver (*.xls, *.xlsx, *.xlsm, *.xlsb)
test	32-bit	Driver do Microsoft Access (*.mdb)

The driver of this User DSN only has a 32-bit version available. It can only be removed or configured with the 32-bit ODBC Data Source Administrator



Create New Data Source

Select a driver for which you want to set up a

Name
Adaptive Server Enterprise
Driver da Microsoft para arquivos texto (*.t
Driver do Microsoft Access (*.mdb)
Driver do Microsoft dBase (*.dbf)
Driver do Microsoft Excel(*.xls)
Driver do Microsoft Paradox (*.db)
Driver para o Microsoft Visual FoxPro

ODBC Microsoft Access Setup

Data Source Name: test

Description: MS Access database for a test

Database

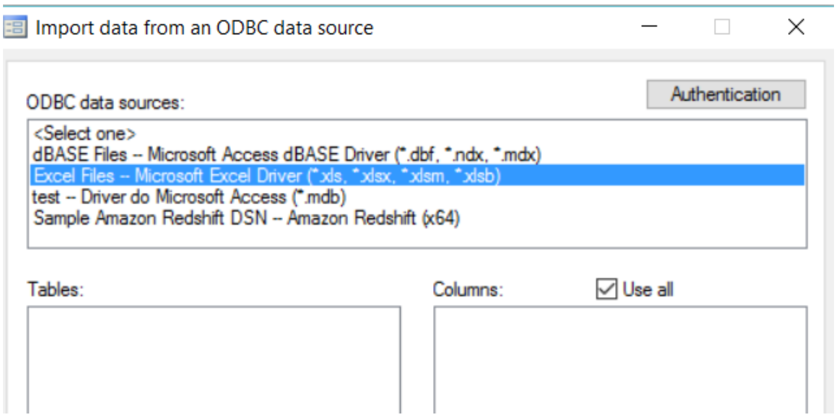
Database:

System Database

● None

○ Database:

In STATA go File=>Import=>ODBC data source. Separate window opens and you can select tables and columns you need.



Import data from an ODBC data source

ODBC data sources:

<Select one>
dBASE Files -- Microsoft Access dBASE Driver (*.dbf, *.ndx, *.mdx)
Excel Files -- Microsoft Excel Driver (*.xls, *.xlsx, *.xlsm, *.xlsb)
test -- Driver do Microsoft Access (*.mdb)
Sample Amazon Redshift DSN -- Amazon Redshift (x64)

Tables:

Columns: ☑ Use all

Additional information for importing data to STATA can be found on this site: https://www.stata.com/meeting/portugal15/abstracts/materials/portugal15_sousa.pdf

Appendix 5 – Overview of pre-selected indicators

Table 1: Overview of pre-selected indicators and example statistics (Cambodia in filter)

category	data_source	subdatabase	indicator_name	short_co	scale_ty	irst_yea	ast_yea	nr_of_ obs	min_value	max_value	average	stdev
BUSENV	World Bank	Doing Business	Extent of shareholder rights index (0-10.5)	shrrin	index	2013	2016	4	1	1	1	0
BUSENV	World Bank	Doing Business	Time required to start a business (days)	trsbd	numeric	2003	2016	14	86	102	96.21	6.81
BUSENV	World Bank	Doing Business	Cost of business start-up procedures (% of GNI per capita)	cbsp	percent	2003	2016	14	57.2	534.8	192.04	147.47
BUSENV	World Bank	Doing Business	Depth of credit information index (0=low to 8=high)	dcinfin	index	2013	2016	4	5	6	5.25	0.5
BUSENV	World Bank	Doing Business	Strength of investor protection index (0 to 10)	strinpin	index	2013	2016	4	4.8	4.8	4.8	0
BUSENV	World Bank	Doing Business	Strength of insolvency framework index (0-16)	strinsfin	index	2013	2016	4	13	13	13	0
BUSENV	World Bank	Doing Business	Minimum paid-in capital required to start a business (% of income p	mincap	percent	2005	2016	12	22.5	80.7	39.6	18.12
BUSENV	World Bank	Doing Business	Start-up procedures to register a business (number)	supregb	numeric	2003	2016	14	8	12	10.57	1.09
BUSENV	World Bank	Doing Business	Extent of director liability index (0 to 10)	dirlibin	index	2005	2016	12	10	10	10	0
BUSENV	World Bank	Doing Business	Profit tax (% of commercial profits)	prftax	percent	2013	2016	4	18.9	19.5	19.35	0.3
BUSENV	World Bank	Doing Business	Total tax rate (% of commercial profits)	ttaxr	percent	2005	2016	12	21	22.6	21.7	0.7
BUSENV	World Bank	Doing Business	Strength of governance structure index (0-10.5)	strgovsir	index	2013	2016	4	3.3	3.3	3.3	0
BUSENV	World Bank	Doing Business	Business extent of disclosure index (0=less disclosure to 10=more d	busdin	index	2005	2016	12	5	5	5	0
BUSENV	World Bank	Doing Business	Extent of corporate transparency index (0-9)	corptrin	index	2013	2016	4	5	5	5	0
BUSENV	World Bank	Doing Business	Ease of doing business index (1=most business-friendly regulations)	edbin		2015	2016	2	128	131	129.5	2.12
BUSENV	World Bank	Doing Business	Other taxes payable by businesses (% of commercial profits)	otaxpb	percent	2014	2016	3	1	1	1	0
BUSENV	World Bank	Doing Business	Labor tax and contributions (% of commercial profits)	ltaxc	percent	2013	2016	4	0.1	0.5	0.4	0.2
BUSENV	World Bank	Doing Business	Strength of legal rights index (0=weak to 12=strong)	srtlrin	index	2013	2016	4	11	11	11	0
BUSENV	World Bank	Doing Business	Extent of conflict of interest regulation index (0-10)	coiregin	index	2013	2016	4	6.3	6.3	6.3	0
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying labor regulations as a major constraint	flabreg	percent	2007	2016	3	1.6	5.2	3.43	1.8
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying the courts system as a major constraint	fcsys	percent	2007	2016	3	4	16.2	10.9	6.26
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying tax rates as a major constraint	ftaxr	percent	2007	2016	3	6.5	23.3	15.37	8.44
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying tax administration as a major constrain	ftaxad	percent	2007	2016	3	4.9	14.8	8.7	5.34
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying practices of competitors in the informal	fpcomp	percent	2007	2016	3	32	36.8	33.87	2.57
BUSENV	World Bank	Enterprise Surveys	If there were losses, average losses due to theft and vandalism (% of	avgltv	percent	2007	2016	3	0.4	3.2	1.37	1.59
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying electricity as a major constraint	fel	percent	2007	2016	3	6.1	33.1	20.07	13.52
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying customs and trade regulations as a maj	fcustr	percent	2007	2016	3	8	13	11.3	2.86
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying crime, theft and disorder as a major con	fcrtd	percent	2007	2016	3	12.3	24.2	18.33	5.95
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying corruption as a major constraint	fcor	percent	2007	2016	3	10.2	53.7	37.23	23.6
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying business licensing and permits as a maj	fbuslp	percent	2007	2016	3	6.1	11.1	8.3	2.55
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying access to finance as a major constraint	ffincon	percent	2007	2016	3	14.2	16.9	15.5	1.35
BUSENV	World Bank	Enterprise Surveys	Percent of firms formally registered when they started operations in	ffreg	percent	2007	2016	3	69.5	87.5	80.17	9.45
BUSENV	World Bank	Enterprise Surveys	If the establishment pays for security, average security costs (% of a	avgsc	percent	2007	2016	3	1	12.4	5.37	6.15
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying an inadequately educated workforce as	finawf	percent	2007	2016	3	15.5	27.3	20.13	6.29
BUSENV	World Bank	Enterprise Surveys	Bribery index (% of gift or informal payment requests during public t	bribin	percent	2007	2016	3	57.8	61.8	59.67	2.01
BUSENV	World Bank	Enterprise Surveys	Percent of firms having their own Web site	fwweb	percent	2007	2016	3	24.2	39.2	33.5	8.12
BUSENV	World Bank	Enterprise Surveys	Proportion of workers offered formal training (%)	fprwot	percent	2013	2013	1	61.3	61.3	61.3	
BUSENV	World Bank	Enterprise Surveys	Proportion of unskilled workers (out of all production workers) (%)	fprusw	percent	2007	2016	3	19.8	48.8	35.77	14.72
BUSENV	World Bank	Enterprise Surveys	Proportion of permanent full-time workers that are female (%)	fprpftf	percent	2007	2016	3	0	46.5	29.47	25.62
BUSENV	World Bank	Enterprise Surveys	Percent of firms with a bank loan/line of credit	fwbc	percent	2007	2016	3	19.9	36.8	25.8	9.53
BUSENV	World Bank	Enterprise Surveys	Percent of firms not needing a loan	fnol	percent	2013	2016	2	58.3	67.1	62.7	6.22
BUSENV	World Bank	Enterprise Surveys	Percent of firms identifying transportation as a major constraint	ftrans	percent	2007	2016	3	9.2	12.9	11.37	1.93
BUSENV	World Bank	World Development In	New businesses registered (number)	nbreg	numeric	2004	2009	6	1049	2826	1966.83	704.78
BUSENV	World Bank	World Development In	Firms with female participation in ownership (% of firms)	fwfpo	percent	2016	2016	1	46.2	46.2	46.2	
BUSENV	World Bank	World Development In	Time spent dealing with the requirements of government regulations	tsgovr	percent	2007	2016	3	1.3	16.4	7.77	7.78
BUSENV	World Bank	World Development In	Firms using banks to finance working capital (% of firms)	fubfwc	percent	2007	2016	3	3.6	18.2	11.47	7.37
BUSENV	World Bank	World Development In	Firms with female top manager (% of firms)	fwftm	percent	2016	2016	1	57.3	57.3	57.3	
DEMO	World Bank	World Development In	Population ages 65 and above (% of total)	pop65_a	percent	1980	2015	36	2.71	4.12	3.2	0.4
DEMO	World Bank	World Development In	Population ages 15-64 (% of total)	pop15_6	percent	1980	2015	36	50.45	64.28	56.68	4.5

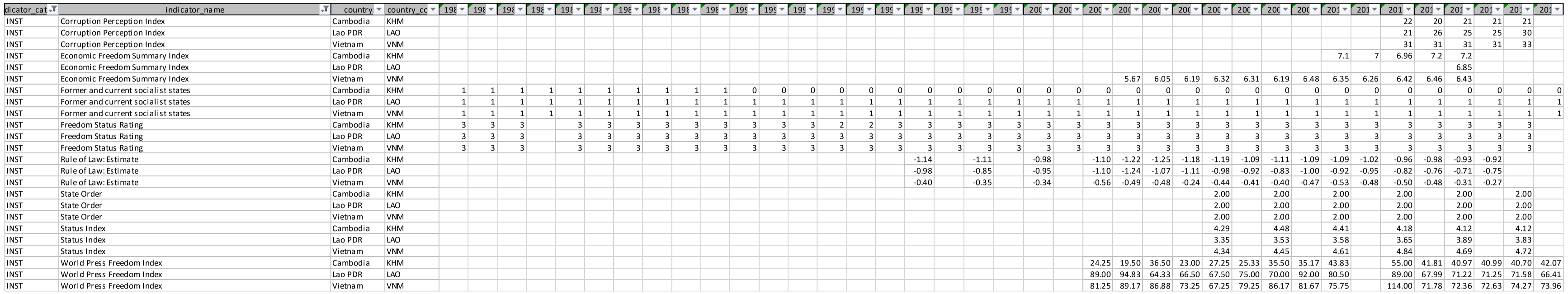
DEMO	World Bank	World Development In	Population, female (% of total)	popf	percent	1980	2015	36	51.22	53.06	51.69	0.52
DEMO	World Bank	World Development In	Population, total	pop	numeric	1980	2015	36	6718241	15577899	11194674.11	2807838.47
DEMO	World Bank	World Development In	Population density (people per sq. km of land area)	popden	numeric	1980	2015	36	38.06	88.25	63.42	15.91
DEMO	World Bank	World Development In	Population in the largest city (% of urban population)	poplcp	percent	1980	2015	36	35.83	53.63	47.33	5.29
DEMO	World Bank	World Development In	Population growth (annual %)	popg	percent	1980	2015	36	-1.14	3.8	2.3	1.02
DEMO	World Bank	World Development In	Population ages 0-14 (% of total)	pop0_14	percent	1980	2015	36	31.6	46.55	40.12	4.85
DEMO	World Bank	World Development In	Population living in slums (% of urban population)	poplsp	percent	2005	2014	2	55.1	78.9	67	16.83
DEMO	World Bank	World Development In	Urban population (% of total)	upop	percent	1980	2015	36	9.9	20.72	17.2	2.72
DEMO	World Bank	World Development In	Rural population (% of total population)	rpop	percent	1980	2015	36	79.28	90.1	82.8	2.72
ECONPER	KOF Swiss Econon		KOF Index of Globalization	kofiin	index10(	1980	2013	34	22.41	50.32	33.42	10.41
ECONPER	World Bank	World Development In	Inflation, consumer prices (annual %)	infcpp	percent	1995	2016	22	-0.8	25	4.91	5.77
ECONPER	World Bank	World Development In	Net ODA received (% of GNI)	nodarp	percent	1995	2015	21	3.97	16.3	8.85	2.93
ECONPER	World Bank	World Development In	Net ODA received per capita (current US$)	nodarpc	numeric	1980	2015	36	1.79	54.45	29.76	18.88
ECONPER	World Bank	World Development In	Renewable electricity output (% of total electricity output)	relop	percent	1995	2014	20	0	61.1	10.44	18.64
ECONPER	World Bank	World Development In	Industry, value added (% of GDP)	indva	percent	1993	2015	23	12.99	29.42	22.51	4.86
ECONPER	World Bank	World Development In	Imports of goods and services (% of GDP)	imgns	percent	1993	2015	23	32.67	76.02	58.98	11.88
ECONPER	World Bank	World Development In	Gross capital formation (% of GDP)	grcapf	percent	1993	2015	23	11.83	22.52	17.69	3.26
ECONPER	World Bank	World Development In	Exports of goods and services (% of GDP)	exgns	percent	1993	2015	23	16.06	68.59	49.84	15.39
ECONPER	World Bank	World Development In	Total debt service (% of exports of goods, services and primary incor	tdsp	percent	1992	2015	24	0.15	9.59	2.31	2.6
ECONPER	World Bank	World Development In	Bank nonperforming loans to total gross loans (%)	bnplp	percent	2010	2016	7	1.59	3.14	2.25	0.54
ECONPER	World Bank	World Development In	Central government debt, total (% of GDP)	cgovd	percent							
ECONPER	World Bank	World Development In	Total reserves (includes gold, current US$)	totres	numeric	1993	2015	23	24181934.5	7306761212	2088253146	2142952082
ECONPER	World Bank	World Development In	Present value of external debt (current US$)	pvexd	numeric	2015	2015	1	4125081315	4125081315	4125081315	
ECONPER	World Bank	World Development In	Deposit interest rate (%)	dintr	percent	1995	2016	22	1.26	8.8	3.5	2.87
ECONPER	World Bank	World Development In	Lending interest rate (%)	lintr	percent							
ECONPER	World Bank	World Development In	Interest rate spread (lending rate minus deposit rate, %)	intrs	percent							
ECONPER	World Bank	World Development In	Real interest rate (%)	rintr	percent							
ECONPER	World Bank	World Development In	Domestic credit to private sector (% of GDP)	dctps	percent	1993	2015	23	2.37	63.1	17.83	17.54
ECONPER	World Bank	World Development In	Grants, excluding technical cooperation (BoP, current US$)	gexctc	numeric	1980	2015	36	5460000	494280000	205383611.1	171970828.8
ECONPER	World Bank	World Development In	Net official development assistance received (current US$)	nodar	numeric	1980	2015	36	13840000	808210000	376847777.8	280013550
ECONPER	World Bank	World Development In	Technical cooperation grants (BoP, current US$)	techco	numeric	1980	2015	36	6500000	196630000	104592222.2	63470152.59
ECONPER	World Bank	World Development In	External debt stocks, private nonguaranteed (PNG) (DOD, current US	exdsppn	numeric	1981	2015	35	0	2440597000	215507428.6	562987248
ECONPER	World Bank	World Development In	Tax revenue (% of GDP)	taxrev	percent	2002	2015	14	7.54	14.56	10.12	2.22
ECONPER	World Bank	World Development In	Bank capital to assets ratio (%)	bctap	percent	2010	2016	7	14.23	20.12	16.19	2.15
ECONPER	World Bank	World Development In	General government final consumption expenditure (% of GDP)	ggfcex	percent	1993	2015	23	3.46	6.93	5.37	0.76
ECONPER	World Bank	World Development In	Income share held by lowest 10%	inclw10	percent	1994	2012	8	2.87	3.93	3.46	0.33
ECONPER	World Bank	World Development In	Income share held by lowest 20%	inclw20	percent	1994	2012	8	6.87	9.05	8.12	0.69
ECONPER	World Bank	World Development In	Agriculture, value added (% of GDP)	agrva	percent	1993	2015	23	28.25	49.62	37.61	6.54
ECONPER	World Bank	World Development In	Services, etc., value added (% of GDP)	serva	percent	1993	2015	23	35.55	42.43	39.87	1.99
ECONPER	World Bank	World Development In	Income share held by highest 10%	inchgh1(	percent	1994	2012	8	25.23	33.57	28.91	2.88
ECONPER	World Bank	World Development In	Income share held by highest 20%	inchgh2(	percent	1994	2012	8	40.21	49.24	43.95	2.93
ECONPER	World Bank	World Development In	Labor force participation rate, female (% of female population ages	lfprtrf	percent	1990	2016	27	73.63	82.27	77.28	2.03
ECONPER	World Bank	World Development In	Gross savings (% of GDP)	gsav	percent	1995	2014	20	4.72	19.37	12.8	4.13
ECONPER	World Bank	World Development In	Household final consumption expenditure, etc. (% of GDP)	hfcexp	percent	1993	2015	23	76.58	100.24	86.08	7.07
ECONPER	World Bank	World Development In	Household final consumption expenditure, etc. (current US$)	hfcex	numeric	1993	2015	23	2539772905	13821976577	6296862643	3724456797
ECONPER	World Bank	World Development In	GDP (current US$)	gdpcur	numeric	1993	2015	23	2533727592	18049954289	7635493371	4980426972
ECONPER	World Bank	World Development In	GNI per capita, PPP (current international $)	gnipcp	numeric	1995	2015	21	790	3300	1784.29	816.99
ECONPER	World Bank	World Development In	Inflation, GDP deflator (annual %)	inf	percent	1994	2015	22	-4.41	12.25	3.58	4.02
ECONPER	World Bank	World Development In	Fuel exports (% of merchandise exports)	fexp	percent	2000	2015	13	0	0.01	0	0
ECONPER	World Bank	World Development In	Personal remittances, received (current US$)	perremr	numeric	1992	2015	24	9000000	397420307.4	133473283.2	101169713.5

ECONPER	World Bank	World Development In	External debt stocks, public and publicly guaranteed (PPG) (DOD, cu	exdsppg	numeric	1981	2015	35	1200000	5419906000	1809728571	1501654835
ECONPER	World Bank	World Development In	GDP growth (annual %)	gdpg	percent	1994	2015	22	0.09	13.25	7.67	2.74
ECONPER	World Bank	World Development In	GDP per capita (current US$)	gdppc	numeric	1993	2015	23	253.19	1158.69	554.72	300.78
ECONPER	World Bank	World Development In	Poverty headcount ratio at $3.10 a day (2011 PPP) (% of population)	phr310d	percent	1994	2012	8	21.58	67.04	38.45	16.39
ECONPER	World Bank	World Development In	Poverty headcount ratio at $1.90 a day (2011 PPP) (% of population)	phr190d	percent	1994	2012	8	2.17	30.06	11.35	9.78
ECONPER	World Bank	World Development In	Employment in industry, male (% of male employment)	epinm	percent	1997	2010	2	5.57	16.97	11.27	8.06
ECONPER	World Bank	World Development In	External debt stocks (% of GNI)	exdsp	percent	1995	2015	21	24.83	60	46.58	10.52
ECONPER	World Bank	World Development In	Employment in industry, female (% of female employment)	epinf	percent	1997	2010	2	3.79	15.52	9.66	8.29
ECONPER	World Bank	World Development In	Foreign direct investment, net inflows (% of GDP)	fdinifp	percent	1993	2015	23	1.75	10.31	6.03	2.77
ECONPER	World Bank	World Development In	Foreign direct investment, net inflows (BoP, current US$)	fdinif	numeric	1992	2015	24	33000000	1730355930	528746199.9	537257608.1
ECONPER	World Bank	World Development In	Foreign direct investment, net (BoP, current US$)	fdinet	numeric	1992	2014	23	-1698435643	-33000000	-467508214.9	474014911.1
ECONPER	World Bank	World Development In	Current account balance (BoP, current US$)	cabal	numeric	1992	2014	23	-1656718571	-87877926.8	-410207995.1	449735211.5
ECONPER	World Bank	World Development In	Foreign direct investment, net outflows (% of GDP)	fdinof	percent	1998	2015	18	-3.44	0.3	-0.29	1.16
ECONPER	World Bank	World Development In	Short-term debt (% of total reserves)	stdp	percent	1993	2015	23	0	51.87	7.37	12.08
ECONPER	World Bank	World Development In	Ores and metals exports (% of merchandise exports)	onmex	percent	2000	2015	16	0	2.79	0.3	0.69
ECONPER	World Bank	World Development In	Children in employment, male (% of male children ages 7-14)	cemm	percent	2001	2012	4	11	52.4	37	18.94
ECONPER	World Bank	World Development In	Income share held by third 20%	incthr20	percent	1994	2012	8	13.79	16.3	15.27	0.84
ECONPER	World Bank	World Development In	Income share held by fourth 20%	incfrt20	percent	1994	2012	8	19.55	21.78	20.88	0.79
ECONPER	World Bank	World Development In	GINI index (World Bank estimate)	giniin	index10(	1994	2012	8	30.76	41.14	35.05	3.36
ECONPER	World Bank	World Development In	Employment in agriculture, female (% of female employment)	eagf	percent	1997	2010	2	55.36	79.87	67.62	17.33
ECONPER	World Bank	World Development In	Employment in agriculture, male (% of male employment)	eagm	percent	1997	2010	2	52.91	73.98	63.45	14.9
ECONPER	World Bank	World Development In	Employment to population ratio, 15+, total (%) (modeled ILO estimat	epopr15t	percent	1991	2016	26	78.4	85.2	80.52	1.62
ECONPER	World Bank	World Development In	Employment in services, female (% of female employment)	emserf	percent	1997	2010	2	14.03	29.07	21.55	10.63
ECONPER	World Bank	World Development In	Income share held by second 20%	incsec20	percent	1994	2012	8	10.29	12.67	11.78	0.79
ECONPER	World Bank	World Development In	Children in employment, female (% of female children ages 7-14)	cemf	percent	2001	2012	4	12.1	52.1	36.55	18.07
ECONPER	World Bank	World Development In	Net migration	netmig	numeric	1982	2012	7	-295987	409414	15698.71	262256.19
ECONPER	World Bank	World Development In	Labor force participation rate, male (% of male population ages 15+	lfprtrm	percent	1990	2016	27	82.32	88.78	85.68	1.69
ECONPER	World Bank	World Development In	Labor force, female (% of total labor force)	labff	percent	1990	2016	27	48.43	51.86	50.2	1
ECONPER	World Bank	World Development In	Labor force, total	labt	numeric	1990	2016	27	4025566	8789877	6383166.07	1600113.95
ECONPER	World Bank	World Development In	Unemployment, youth female (% of female labor force ages 15-24) (r	uemyf	percent	1991	2016	26	0.12	2.76	1.04	0.81
ECONPER	World Bank	World Development In	Unemployment, youth male (% of male labor force ages 15-24) (mod	uemym	percent	1991	2016	26	0.21	5.01	1.9	1.48
ECONPER	World Bank	World Development In	Unemployment, female (% of female labor force) (modeled ILO estim	uemf	percent	1991	2016	26	0.08	1.96	0.74	0.59
ECONPER	World Bank	World Development In	Unemployment, male (% of male labor force) (modeled ILO estimate)	uemm	percent	1991	2016	26	0.12	3.06	1.17	0.93
ECONPER	World Bank	World Development In	International tourism, expenditures (% of total imports)	inttex	percent	1995	2014	20	1.6	4.22	2.94	0.87
ECONPER	World Bank	World Development In	Employment in services, male (% of male employment)	emserm	percent	1997	2010	2	18.35	30.06	24.2	8.28
ECONPER	World Bank	World Development In	GNI, PPP (current international $)	gnip	numeric	1995	2015	21	8415663686	51397159601	24742239727	13627142446
ECONPER	World Bank	World Development In	GDP per capita growth (annual %)	gdppcg	percent	1994	2015	22	-1.4	11.48	5.53	2.83
ECONPER	World Bank	World Development In	Official exchange rate (LCU per US$, period average)	ofexra	numeric	1990	2016	27	426.25	4184.92	3399.65	1087.64
ECONPER	World Bank	World Development In	GDP per capita, PPP (current international $)	gdppcp	numeric	1993	2015	23	707.08	3490.42	1766.59	894.63
ECONPER	World Bank	World Development In	International tourism, receipts (% of total exports)	inttr	percent	1995	2014	20	7.33	30.18	21.51	6.72
ECONPER	World Bank	World Development In	Real effective exchange rate index (2010 = 100)	reexrin	numeric							
EDU	World Bank	Education statistics	PISA: Mean performance on the science scale. Female	pisasf	numeric							
EDU	World Bank	Education statistics	PISA: Mean performance on the science scale	pisas	numeric							
EDU	World Bank	Education statistics	PISA: Mean performance on the reading scale. Male	pisarm	numeric							
EDU	World Bank	Education statistics	DHS: Gross attendance rate. Post Secondary. Male	garpsm	percent	2000	2014	4	2.81	9.31	5.82	3.38
EDU	World Bank	Education statistics	DHS: Gross attendance rate. Post Secondary. Rural	garpsr	percent	2000	2014	4	0.56	3.98	2.35	1.81
EDU	World Bank	Education statistics	DHS: Gross attendance rate. Post Secondary. Female	garpsf	percent	2000	2014	4	1.57	7	4.36	3.01
EDU	World Bank	Education statistics	PISA: Mean performance on the mathematics scale. Female	pisamf	numeric							
EDU	World Bank	Education statistics	DHS: Gross attendance rate. Post Secondary	garps	percent	2000	2014	4	2.19	8.16	5.09	3.19
EDU	World Bank	Education statistics	PISA: Mean performance on the science scale. Male	pisasm	numeric							

EDU	World Bank	Education statistics	DHS: Gross attendance rate. Post Secondary. Urban	garpsu	percent	2000	2014	4	8.67	23.22	16.2	7.89
EDU	World Bank	Education statistics	PISA: Mean performance on the mathematics scale	pisam	numeric							
EDU	World Bank	Education statistics	PISA: Mean performance on the reading scale. Female	pisarf	numeric							
EDU	World Bank	Education statistics	PISA: Mean performance on the reading scale	pisar	numeric							
EDU	World Bank	Education statistics	PISA: Mean performance on the mathematics scale. Male	pisamm	numeric							
EDU	World Bank	World Development In	Over-age students, primary (% of enrollment)	oastp	percent	1998	2015	12	0	22.1	12.05	9.29
EDU	World Bank	World Development In	School enrollment, tertiary (% gross)	schent	percent	1980	2015	29	0.08	15.9	3.59	4.67
EDU	World Bank	World Development In	Government expenditure per student, secondary (% of GDP per capita	govexpss	percent	1998	2001	2	6.08	10.87	8.48	3.39
EDU	World Bank	World Development In	Over-age students, primary, male (% of male enrollment)	oastpm	percent	1998	2015	12	0	23.22	12.66	9.72
EDU	World Bank	World Development In	Literacy rate, adult male (% of males ages 15 and above)	litram	percent	1998	2015	5	79.48	85.08	83.39	2.38
EDU	World Bank	World Development In	Government expenditure per student, tertiary (% of GDP per capita)	govexpst	percent	2001	2011	4	5.27	41.81	15.63	17.51
EDU	World Bank	World Development In	Government expenditure on education, total (% of government expen	govexet	percent	1998	2014	12	7.54	12.4	9.65	1.63
EDU	World Bank	World Development In	Government expenditure on education, total (% of GDP)	govexetp	percent	1998	2014	12	1.26	2.02	1.66	0.19
EDU	World Bank	World Development In	School enrollment, secondary (% gross)	schens	percent	1991	2008	11	16.55	45.05	27.91	10.51
EDU	World Bank	World Development In	School enrollment, primary (% gross)	schenp	percent	1981	2015	32	92.87	211.3	134.25	34.6
EDU	World Bank	World Development In	Over-age students, primary, female (% of female enrollment)	oastpf	percent	1998	2015	12	0	21.19	11.38	8.82
EDU	World Bank	World Development In	Literacy rate, adult female (% of females ages 15 and above)	litraf	percent	1998	2015	5	56.99	72.3	66.03	6.09
EDU	World Bank	World Development In	Literacy rate, youth male (% of males ages 15-24)	litrym	percent	1998	2015	5	81.85	91.12	87.73	3.51
EDU	World Bank	World Development In	Literacy rate, adult total (% of people ages 15 and above)	litrat	percent	1998	2015	5	67.34	78.35	74.16	4.37
EDU	World Bank	World Development In	Literacy rate, youth female (% of females ages 15-24)	litryf	percent	1998	2015	5	71.07	91.97	82.67	7.96
EDU	World Bank	World Development In	Government expenditure per student, primary (% of GDP per capita)	govexpsp	percent	1998	2014	10	4.84	6.81	5.61	0.77
EDU	World Bank	World Development In	Pupil-teacher ratio, primary	ptrp	numeric	1981	2015	32	32.57	56.29	45.52	6.5
HEALTH	World Bank	World Development In	Birth rate, crude (per 1,000 people)	birrcr	numeric	1980	2015	36	23.78	50.18	34.53	9.63
HEALTH	World Bank	World Development In	Fertility rate, total (births per woman)	ferrt	numeric	1980	2015	36	2.6	6.34	4.4	1.34
HEALTH	World Bank	World Development In	Health expenditure per capita (current US$)	hexpc	numeric	1995	2014	20	16.19	61.28	32.01	16.12
HEALTH	World Bank	World Development In	Life expectancy at birth, female (years)	lexpf	numeric	1980	2015	36	30.49	70.75	58.81	9.39
HEALTH	World Bank	World Development In	Health expenditure, public (% of total health expenditure)	hexp	percent	1995	2014	20	20.01	37.19	26.26	5.7
HEALTH	World Bank	World Development In	Health expenditure, total (% of GDP)	hext	percent	1995	2014	20	3.75	7.43	5.85	0.77
HEALTH	World Bank	World Development In	Life expectancy at birth, male (years)	lexpm	numeric	1980	2015	36	25.12	66.66	54.28	9.59
HEALTH	World Bank	World Development In	Life expectancy at birth, total (years)	lexpt	numeric	1980	2015	36	27.74	68.66	56.49	9.49
ICTINFRA	World Bank	Education statistics	Personal computers (per 100 people)	pcomp	numeric	1995	2007	13	0.05	0.38	0.19	0.12
ICTINFRA	World Bank	World Development In	Fixed telephone subscriptions (per 100 people)	telsub	numeric	1987	2015	29	0.03	3.93	0.71	1.15
ICTINFRA	World Bank	World Development In	Internet users (per 100 people)	intus	numeric	1990	2015	20	0	19	2.62	5.15
ICTINFRA	World Bank	World Development In	Mobile cellular subscriptions (per 100 people)	mobsub	numeric	1980	2015	36	0	133.89	22.54	43.78
ICTINFRA	World Bank	World Development In	Fixed broadband subscriptions (per 100 people)	brsub	numeric	2002	2015	14	0	0.53	0.16	0.17
ICTINFRA	World Bank	World Development In	Rail lines (total route-km)	railln	numeric	1990	2005	14	600	650	604.07	13.26
INNOSYS	World Bank	World Development In	Researchers in R&D (per million people)	rirnd	numeric	2002	2002	1	17.58	17.58	17.58	
INNOSYS	World Bank	World Development In	Trademark applications, direct nonresident	tapdnr	numeric	1994	2015	22	548	4886	1968.73	1215.97
INNOSYS	World Bank	World Development In	Charges for the use of intellectual property, payments (BoP, current	chintpp	numeric	1998	2014	17	4435554.2	20506256.61	8086330.92	3934253.95
INNOSYS	World Bank	World Development In	Charges for the use of intellectual property, receipts (BoP, current U	chintpr	numeric	2003	2014	12	19756.5	3840000	986854.22	1265866.88
INNOSYS	World Bank	World Development In	Technicians in R&D (per million people)	tirnd	numeric	2002	2002	1	13.37	13.37	13.37	
INNOSYS	World Bank	World Development In	Scientific and technical journal articles	sntja	numeric	1986	2013	28	0	84	23.83	27.33
INNOSYS	World Bank	World Development In	Patent applications, residents	papr	numeric	2013	2014	2	1	2	1.5	0.71
INNOSYS	World Bank	World Development In	Trademark applications, direct resident	tapdr	numeric	1994	2014	21	3	1182	467.81	377.56
INNOSYS	World Bank	World Development In	High-technology exports (% of manufactured exports)	htexp	percent	2000	2015	16	0.03	0.76	0.19	0.19
INNOSYS	World Bank	World Development In	High-technology exports (current US$)	htex	numeric	2000	2015	16	957962	60108587	9839339.5	15345861.96
INNOSYS	World Bank	World Development In	Research and development expenditure (% of GDP)	rndex	percent	2002	2002	1	0.05	0.05	0.05	
INNOSYS	World Bank	World Development In	Patent applications, nonresidents	papnr	numeric	2007	2015	9	13	74	45.11	20.67
INST	Transparency Inte		Corruption Perception Index	cpin	index	2012	2016	5	20	22	21	0.71
INST	Freedom House		Freedom Status Rating	frstrat	nominal	1980	2016	36	2	3	2.94	0.23

INST	Bertelsmann Foun		Status Index	btistin	index	2006	2016	6	4.12	4.48	4.26	0.15
INST	CIA factbook		Former and current socialist states	issoc	nominal	1980	2017	38	0	1	0.26	0.45
INST	Bertelsmann Foun		State Order	stord	binary	2006	2016	6	2	2	2	0
INST	Fraser Institute		5 Regulation	efsin5	index	2010	2014	5	6.04	7.13	6.47	0.45
INST	Freedom House		Civil Liberties Rating	fsrclin	index	1980	2016	36	5	7	5.78	0.87
INST	Freedom House		Political Rights Rating	fsprin	index	1980	2016	36	4	7	6.19	0.71
INST	Fraser Institute		Economic Freedom Summary Index	efsin	index	2010	2014	5	6.96	7.2	7.09	0.11
INST	Fraser Institute		1 Size of Government	efsin1	index	2010	2014	5	7.85	7.88	7.87	0.01
INST	Fraser Institute		4 Freedom to trade internationally	efsin4	index	2010	2014	5	7.03	7.88	7.37	0.35
INST	Fraser Institute		3 Sound Money	efsin3	index	2010	2014	5	9.25	9.5	9.31	0.11
INST	Fraser Institute		2 Legal System & Property Rights	efsin2	index	2007	2014	6	4.1	4.76	4.46	0.26
INST	Reporters Withou		World Press Freedom Index	prfrin	index10	2002	2017	15	19.5	55	35.46	9.75
INST	Bertelsmann Foun		Management Index	btimgtin	index	2006	2016	6	3.4	3.82	3.62	0.17
INST	World Bank	Country Policy and Ins	CPIA economic management cluster average (1=low to 6=high)	cpiaem	index	2005	2015	11	3.5	4	3.8	0.15
INST	World Bank	Country Policy and Ins	CPIA structural policies cluster average (1=low to 6=high)	cpiasp	index	2005	2015	11	3	3.67	3.39	0.23
INST	World Bank	Country Policy and Ins	CPIA policies for social inclusion/equity cluster average (1=low to 6	cpiapsi	index	2005	2015	11	3.1	3.5	3.35	0.11
INST	World Bank	Country Policy and Ins	CPIA public sector management and institutions cluster average (1=	cpiapsm	index	2005	2015	11	2.6	2.8	2.73	0.06
INST	World Bank	World Development In	Overall level of statistical capacity (scale 0 - 100)	ovlsc	index10	2004	2016	13	64.44	76.67	71.28	3.79
INST	World Bank	World Development In	Intentional homicides (per 100,000 people)	inth	numeric	1995	2011	16	1.8	6.76	3.75	1.4
INST	World Bank	World Development In	Female legislators, senior officials and managers (% of total)	flom	percent							
INST	World Bank	World Governance Ind	Voice and Accountability: Estimate	vacc	numeric	1996	2015	17	-1.1	-0.78	-0.94	0.08
INST	World Bank	World Governance Ind	Regulatory Quality: Estimate	regq	numeric	1996	2015	17	-0.58	-0.05	-0.39	0.15
INST	World Bank	World Governance Ind	Rule of Law: Estimate	rol	numeric	1996	2015	17	-1.25	-0.92	-1.08	0.1
INST	World Bank	World Governance Ind	Political Stability and Absence of Violence/Terrorism: Estimate	psav	numeric	1996	2015	17	-1.3	-0.03	-0.51	0.36
INST	World Bank	World Governance Ind	Government Effectiveness: Estimate	goveff	numeric	1996	2015	17	-1.07	-0.68	-0.88	0.09
INST	World Bank	World Governance Ind	Control of Corruption: Estimate	cocor	numeric	1996	2015	17	-1.23	-0.85	-1.08	0.11
STRUCT	World Bank	World Development In	Arable land (% of land area)	arblap	percent	1980	2014	35	11.33	21.53	19.12	3.66
STRUCT	World Bank	World Development In	Access to electricity (% of population)	actelp	percent	1991	2014	24	0.1	56.1	21.01	15.01
STRUCT	World Bank	World Development In	Forest area (% of land area)	forarp	percent	1990	2015	26	53.57	73.33	63.25	6.16
STRUCT	World Bank	World Development In	Electric power consumption (kWh per capita)	elpcon	numeric	1995	2014	20	13.46	270.42	91.75	75.22
STRUCT	World Bank	World Development In	Surface area (sq. km)	surfar	numeric	1980	2016	37	181040	181040	181040	0
STRUCT	World Bank	World Development In	Land area (sq. km)	landar	numeric	1980	2016	37	176520	176520	176520	0

Table 2: Timeline overview of select institutional indicators for Cambodia, Laos and Vietnam (example)



dicator_cat	indicator_name	country	country_cc	198	198	198	198	198	198	198	198	198	198	199	199	199	199	199	199	199	199	199	199	200	200	200	200	200	200	200	200	200	200	201	201	201	201	201	201	201	201
INST	Corruption Perception Index	Cambodia	KHM																																	22	20	21	21	21	
INST	Corruption Perception Index	Lao PDR	LAO																																	21	26	25	25	30	
INST	Corruption Perception Index	Vietnam	VNM																																	31	31	31	31	33	
INST	Economic Freedom Summary Index	Cambodia	KHM																															7.1	7	6.96	7.2	7.2			
INST	Economic Freedom Summary Index	Lao PDR	LAO																																			6.85			
INST	Economic Freedom Summary Index	Vietnam	VNM																							5.67	6.05	6.19	6.32	6.31	6.19	6.48	6.35	6.26	6.42	6.46	6.43			
INST	Former and current socialist states	Cambodia	KHM	1	1	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
INST	Former and current socialist states	Lao PDR	LAO	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
INST	Former and current socialist states	Vietnam	VNM	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
INST	Freedom Status Rating	Cambodia	KHM	3	3	3		3	3	3	3	3	3	3	3	3	2	2	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
INST	Freedom Status Rating	Lao PDR	LAO	3	3	3		3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
INST	Freedom Status Rating	Vietnam	VNM	3	3	3		3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	3	
INST	Rule of Law: Estimate	Cambodia	KHM																	-1.14		-1.11		-0.98		-1.10	-1.22	-1.25	-1.18	-1.19	-1.09	-1.11	-1.09	-1.09	-1.02	-0.96	-0.98	-0.93	-0.92		
INST	Rule of Law: Estimate	Lao PDR	LAO																	-0.98		-0.85		-0.95		-1.10	-1.24	-1.07	-1.11	-0.98	-0.92	-0.83	-1.00	-0.92	-0.95	-0.82	-0.76	-0.71	-0.75		
INST	Rule of Law: Estimate	Vietnam	VNM																	-0.40		-0.35		-0.34		-0.56	-0.49	-0.48	-0.24	-0.44	-0.41	-0.40	-0.47	-0.53	-0.48	-0.50	-0.48	-0.31	-0.27		
INST	State Order	Cambodia	KHM																											2.00		2.00		2.00		2.00		2.00		2.00	
INST	State Order	Lao PDR	LAO																											2.00		2.00		2.00		2.00		2.00		2.00	
INST	State Order	Vietnam	VNM																											2.00		2.00		2.00		2.00		2.00		2.00	
INST	Status Index	Cambodia	KHM																											4.29		4.48		4.41		4.18		4.12		4.12	
INST	Status Index	Lao PDR	LAO																											3.35		3.53		3.58		3.65		3.89		3.83	
INST	Status Index	Vietnam	VNM																											4.34		4.45		4.61		4.84		4.69		4.72	
INST	World Press Freedom Index	Cambodia	KHM																							24.25	19.50	36.50	23.00	27.25	25.33	35.50	35.17	43.83		55.00	41.81	40.97	40.99	40.70	42.07
INST	World Press Freedom Index	Lao PDR	LAO																							89.00	94.83	64.33	66.50	67.50	75.00	70.00	92.00	80.50		89.00	67.99	71.22	71.25	71.58	66.41
INST	World Press Freedom Index	Vietnam	VNM																							81.25	89.17	86.88	73.25	67.25	79.25	86.17	81.67	75.75		114.00	71.78	72.36	72.63	74.27	73.96