Contact
CaleaDorobanţilor, no. 15-17, room 2017, Bucharest, Romania
Web: http://dbjournal.ro
E-mail: editordbjournal@gmail.com; editor@dbjournal.ro

CONTENTS

Data model for Demand Side Management

Simona-Vasilica OPREA[1], Osman Bulent TOR[2]
[1]The Bucharest university of Economic Studies, Romania
[2]EPRA Engineering Procurement Research Analysis, Turkey
simona.oprea@csie.ase.ro;osman.tor@epra.com.tr

Demand Side Management (DSM) is a portfolio of measures to improve the energy system mainly at the consumption level. In this paper we propose a data model for DSM stating from the optimization methods approach in SMARTRADE project from different perspectives of several entities that include: Transmission System Operator (TSO)/Distribution System Operators (DSOs) perspectives in case of security/reliability concerns: minimum amount of load (or generation) shedding; aggregators perspective in case of demand or generation shedding request: Which demand (or generators) should be shed?; consumers perspective: load shifting (time-of-use (ToU) tariffs) and optimum contract strategies with the aggregators (also known as balancing responsible parties- BRP) for load shedding.
Keywords: *data model, demand side management, optimization process, mixed integer linear programing, electricity consumption, load/generation shedding*

1 Introduction

DSM ranges from improving energy efficiency by using energy efficient products, over smart energy tariffs with incentives for certain consumption patterns, up to sophisticated real-time control of load and distributed energy resources [1]. DSM is an important function in energy management of the future smart grid, which provides support towards smart grid functionalities in various areas such as electricity market control and management, infrastructure, management of decentralized energy resources, etc. Controlling and influencing energy demand can reduce the overall peak load demand, reshape the demand profile and increase the grid sustainability by reducing the overall cost and carbon emission levels.

DSM is becoming more and more important in smart grids concept that allows consumers to make informed decisions regarding their energy consumption and helps the energy suppliers reduce the peak load demand and reshape the load profile [2]. DSM models for different time scales are presented in Fig. 1. For long-term perspective, the DSM provides improvement in energy efficiency; it enables direct load control in real time when necessary for the security and/or reliability of the power system.

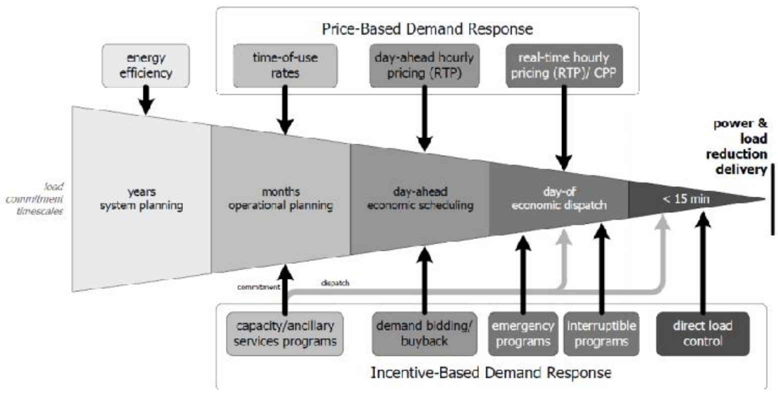

Fig. 1. DSM models for different time scales

Time-of-use tariff rates are among the initial implementation of DSM, which has been implemented in many countries for several years. Time-of-use tariff rates enable consumers who are willing to reduce consumption during periods when the total demand for electricity is highest (at peak load) to save money. Consumers who are subject to time-of-use rates can reduce their expenses by shifting their energy use to partial-peak or off-peak hours of the day. In order to benefit from time-of-use tariffrates, the electricity meters of the consumers should be electronic and have the capability to stamp time according to rate periods (e.g., morning time, peak time, night time, etc.). Smart meters nowadays provide more than stamping time of use indeed. They provide messages from electricity suppliers, reading remotely and direct load control (electricity cut) via a telemetry. Smart meters are necessary for all types of price-based and incentive-based demand responses as illustrated in Fig. 1.

2 SMARTRADE Project in terms of DSM and electricity market

The research project "Intelligent system for trading on wholesale electricity market" (with acronym SMARTRADE) is supported by National Authority for Scientific Research and Innovation through European Regional Development Fund (ERDF). The main objective of the project is to design and develop an informatics prototype for forecasting, analysis and decision models mainly for all interested market participants (suppliers/producers) constituted as balance responsible parties (BRP) or aggregators, in order to estimate energy demand and generation in a suitable way for an efficient trading on the wholesale energy market. The prototype will be developed on a private cloud computing architecture and will be addressed to BRP and grid operators alike, especially to the Transmission System Operator and the Distribution System Operators, for estimation of the electricity demand and generation at the national or regional level. An important component of the informatics prototype consists in a forecast module that accurately predict the electricity generation/demand on short and medium term. Main scope of the project is to establish efficient trading offers on the energy market, based on business rules and decision models.

First stage of the project is the conceptual design including: resources, methodologies and technologies that can be defined and used in the project. Conceptual design of the prototype is made based on following main objective: to develop a software platform which will be utilized to ensure supply and demand balance along a planning period in the electricity market. Supply and demand balance should satisfy transmission grid constraints in an optimum manner. Re-dispatching the generators based on merit order, cutting of generation from renewables when they are generating quite high amount of energy and load shedding are among those solutions to ensure supply and demand balance while satisfying grid constraints [3]-[5]. Whether the solution is optimum or not in terms of cost effectiveness is the main question, which is intended to be answered by means of the proposed optimization techniques in the software platform. Modules of the SMARTRADE project prototype are presented in Fig. 2 as described in [7].

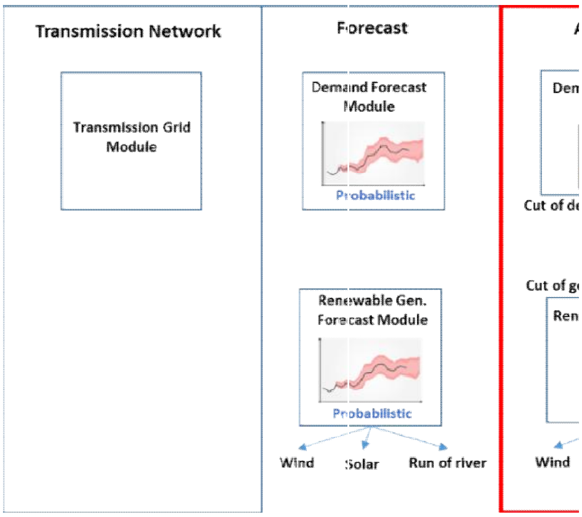

Fig. 2. Modules of the SMARTRADE project prototype [7]

The key component of the prototype is the aggregator module, as marked in Fig. 2. It is considered in modelling either demand or renewable generation shedding, as necessary. Demand and renewable generation aggregator moduleconsists in representing demand and renewable generation shedding capabilities for supply and demand balancing issue.

Aggregators are becoming one of the key players in electricity market along with other developments in the power sector including: maturation of the electricity markets; increase of distributed generation which are embedded in distribution grid (wind, PV, biomass and other renewable sources); smart meters implementation; transformation of some consumers into prosumers.

Initial electricity market structure in many countries worldwide include a wholesale market at high voltage level (i.e., transmission), as shown in Fig. 3. Given the developments described above, aggregators has started to take roles in this market by providing load or generation shedding services to TSO, as shown in Fig. 4. However, this will also introduce some complexities to the DSOs, particularly when demand (or generation) shedding is required at the distribution grid level. Main question for the DSO is: how would the demand or generation shedding affect the distribution system? For instance, if demand shedding is made on a distribution feeder which has several PV units generating power, voltage on the feeder may increase to unacceptable levels.

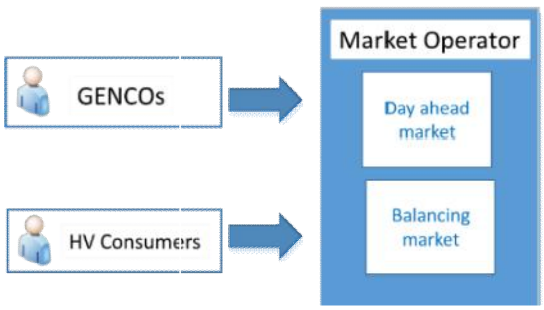

Fig. 3. Wholesale market at transmission level

Therefore, DSOs should essentially be included in the final decision of demand or generator shedding. Actually, there are discussions to develop a balancing market at the distribution grid level as well, in which aggregators can provide servicesdirectly to DSOs. Such a mechanism require coordination between the market operator, TSO, DSOs, aggregators and other players, as illustrated in Fig. 5. In conclusion, aggregators are expected to be one of the key players which provide demand and generation shedding services to TSO and DSOs in the future electricity markets. Therefore, aggregators are considered as a separate module in the SMARTRADE project prototype.

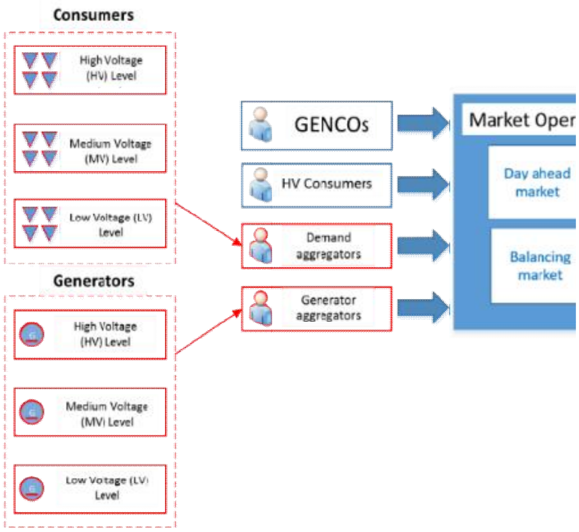

Fig. 4. Wholesale market at transmission grid level

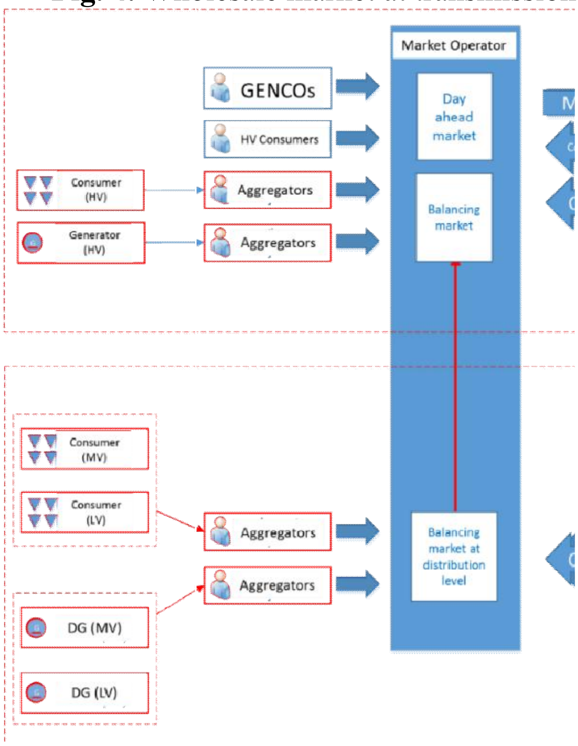

Fig. 5. Balancing market at the distribution grid level

3 Optimization methods in DSM process and proposed approach in SMARTRADE project

DSM concept is an optimization process inherently. Therefore, the optimization processes from the perspectives of market players include:

- From TSO/DSOs perspectives → In case of security/reliability concerns:
  - Minimum amount of load (or generation) shedding;

- From aggregators perspective → In case of demand or generation shedding request:
  - Which demand (or generators) should be shed?;
- From consumers perspective:
  - Load shifting (time-of-use tariffs);
  - Optimum contract strategies with the aggregators (load shedding).

Essentially, each market player should develop their optimization methods according to their objective function.

In the SMARTRADE project, perspectives of both TSO/DSOs and aggregators will be considered. Also SMARTRADE project will address optimizing load (or generation) shedding from TSO perspective, as illustrated in Fig. 4. The word not "minimizing", but "optimizing" of shedding is utilized due to fact that shedding decision is based on optimizing the load shedding (or generator) bids of the aggregators taking into account constraint of transmission system. If a TSO requires load or generation shedding to relax the power system which is subjected to network constraints, it should optimize this request based on cost of load or generation shedding. Bids from the aggregators are basis for determining the shedding cost.

Load or generator shedding could be considered in both operation and planning problem. In the planning problem (from day-ahead planning to long-term investment planning), market simulation gives commitment of the generators based on merit order, when the network constraints are ignored [6]. Objective function of the market simulation is minimizing total cost of generation. Generation of renewables, like wind power plants, PVs, and run-of-river hydraulic power plants are assumed as negative generation. A corresponding cost is defined for shedding the generation from renewables, in case of necessity due to technical constraints.

Technical constraints could be: i) overloading on the transmission branches and ii) N-1 contingency criterion (the rule according to which after one network element (line, transformer) failure, the elements remaining in operation within TSO's responsibility area after a contingency from the contingency list must be capable of accommodating the new operational situation without violating operational security limits).

Shedding cost of renewables is determined from the bids of aggregators who are assumed to have contracts with renewable generation sources and have capability of shedding the load (i.e., already equipped with telemetry infrastructure, etc.). If total generation is higher than load at certain hours, market simulation essentially proposes shedding the renewable generation starting from the minimum cost, if the cost of shedding generation from renewable sources is higher than that of conventional sources.

Market simulation result is given as input to the network simulation by the TSO, as illustrated in Fig. 6. Network simulation re-dispatches the conventional generators from their initial dispatch levels proposed by the market simulation results, based on either transmission constraints or spinning reserve requirements or both. Costs of generator and load shedding is considered in the network simulations as well as market simulation. Network simulation will determine optimum solution considering network constraints and necessary re-dispatches and/or load and/or generator shedding.

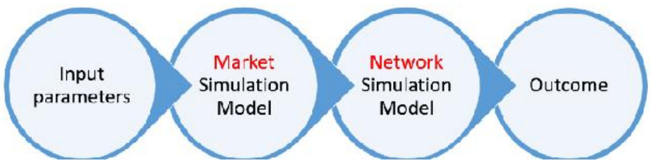

Fig. 6. Sequence simulation approach (market simulation → network simulation)

Both market simulation and network simulation problems are dynamic optimization problems along a time interval (e.g., 24 hour of a day for day-ahead planning or 8760 hour of a year for one year planning). They are both mixed-integer linear programming (MILP) optimization problems.

Decomposition techniques will be utilized to get optimum solution iteratively, as illustrated in Fig. 7. Conceptual design of the proposed solution (market simulation → network simulation) is presented in details in **Error! Reference source not found.** . It may be assumed that there is no need for generator and/or load shedding in the market simulation. However, in case of total generation of renewables is higher than total load at certain hours, bids of aggregators could essentially be considered in market simulation as well. Similarly, if the total capacity of generation resources including renewables are less than total load at certain hours, load shedding could be considered based on bids of aggregators.

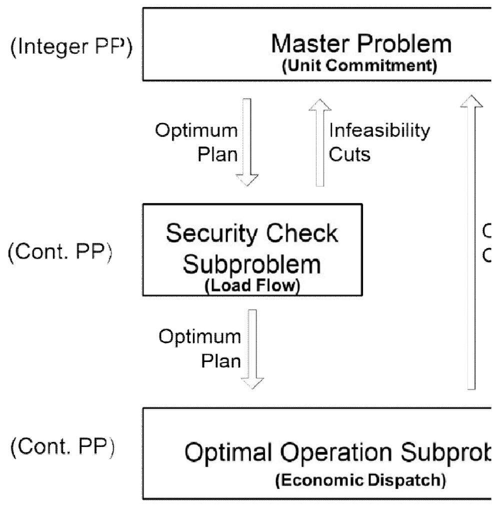

Fig. 7. Decomposition technique for iterative calculation of the optimum solution in network simulation

Load or generator shedding bids of the aggregators should be determined based on strategic decisions of the aggregators. This requires aggregators to define optimization processes. The aggregator bids, which are input to the market and network simulations of TSOare assumed to be identified by the aggregators in these modules.

Handling storage devices plays an important role in DSM; they can be categorized mainly as follows:

- Hydraulic power plants (HPP) which has storage dam:
    - Only generation;
- Pump storage hydraulic power plants:
    - Generation and consumption;
- Batteries:
    - Generation and consumption;
- Other storage devices (hydrogen, etc.).

All types of storages are subjected to energy constraints. For example, an HPP which has a large storage dam has a long-term energy constraint. That is, its energy production level is limited by its capacity stored in the dam. For pump storage power plants, energy constraint is rather on short-term horizon. Energy production cycle may be less than a day depending on the physical dimension of its storage, as illustrated by an example in Fig. [78]. Energy constraints of the batteries and other storage devices are even shorter in terms of duration.

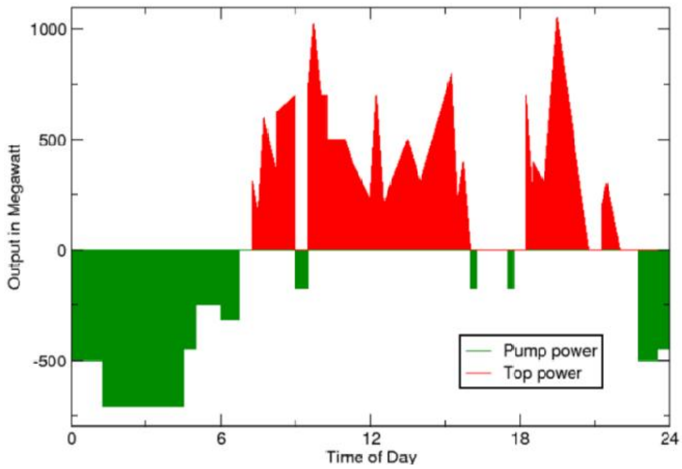

Fig. 8. Generating and pumping cycle of a pump storage power plant

In the SMARTRADE project, large dam HPPs and pump storages are assumed individual players in the market. However, battery and other storage devices are assumed to play in the market through the aggregators.

Cyclic efficiency of the batteries will be considered in the model. Efficiency of the battery devices will be assumed to be constant for short-term simulations, like one day, one week, etc. However, efficiency evolution of batteriesdue to charging and discharging along their lifetime will be considered for long-term simulations (one year and beyond).

DSM will be assumed to be provided by aggregators which will be modelled by demand aggregator module and generation aggregator module, as presented in [7]. Main data inputs to these modules include: aggregated time series demand, which can be shed by aggregator upon request from the TSO; time series bids of aggregators for demand shedding; aggregated time series generation, which can be shed by aggregator upon request from the TSO; time series bids of aggregators for generator shedding; aggregated storage (battery) capacity; energy constraints and cyclic efficiency of the batteries; time series bids of aggregators for storage charging and discharging periods will be utilized for demand and generator shedding bids, respectively.

4 Data model for DSM

First stage of the data model for DSM presented in Fig. 9 contains the following main tables: T_ELECTRICITY_SUPPLIER withdata regarding the electricity supplier; T_SUPPLIER_MARKET withdata regarding wholesale electricity markets' participation of the supplier; T_ PROFILES for consumers' profiles determined by the electricity supplier; T_CONSUMER with data regarding consumers; T_CONSUMER_PLACE depicting place for consumption with mention that each consumer may have one or multiple consumption places; T_CONSUMER_PLACE_DETAILS showing consumer's place detailed information regarding occupancy and generation sources; T_TARIFFS for electricity tariffs' schemas applied by electricity supplier to their consumers; T_CONSUMER_TARIFFS with electricity tariffs' schemas applied to a specific consumer; T_METER for smart meters installed at consumer's side; T_METER_READINGS for smart meters reading transmitted to the electricity supplier; T_ELECTRIC_APPLIANCES for electric appliances installed at consumer's side; T_APPLIANCES_TYPES for electric appliances' detailed information; T_ APPLIANCE_READING for electric appliances' reading transmitted to the electricity supplier.

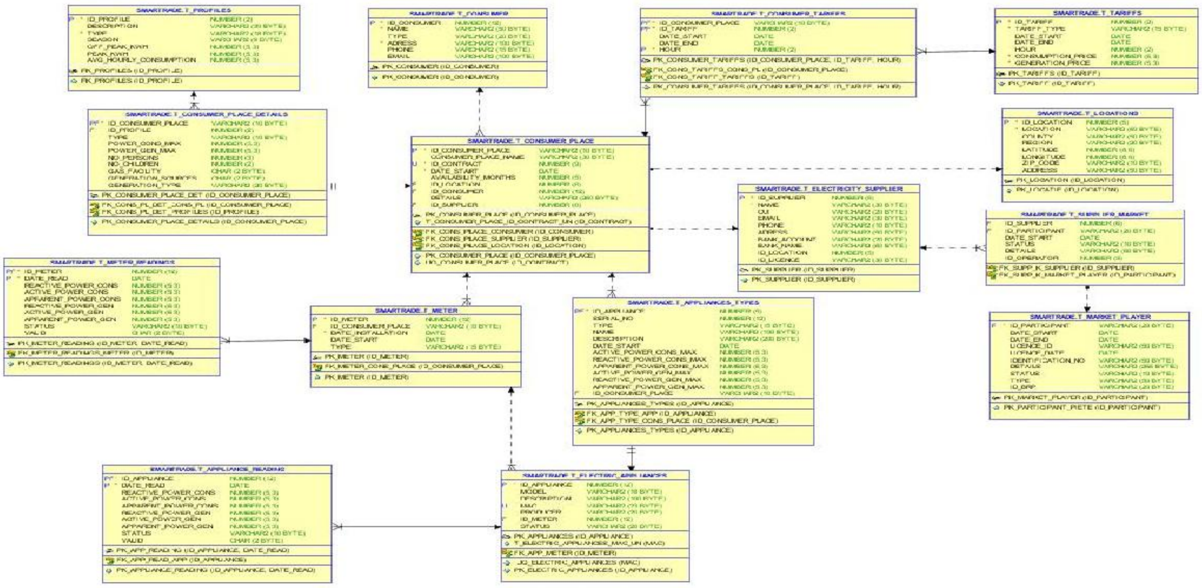

Fig. 9. Data model for DSM

The data model allow the monitoring and achieving of specific activities for DSM measures. However, it will be continuously refined by inclusion of other measures and concepts (i.e. IoT).

5 Conclusions

In this paper proposed a data model for DSM stating from the optimization methods approach in SMARTRADE project from different perspectives of several entities that include: TSO/DSOs, aggregators, consumers' perspective. Data model contains tables related to the electricity supplier; wholesale electricity markets' participation of the supplier; consumers' profiles determined by the electricity supplier; data regarding consumers; depicting place for consumption; consumers' place detailed information regarding occupancy and generation sources; electricity tariffs' schemas applied by electricity supplier to their consumers; electricity tariffs' schemas applied to a specific consumer; smart meters data installed at consumer's side; smart meters reading transmitted to the electricity supplier; electric appliances installed at consumer's side; electric appliances' detailed information; electric appliances' reading transmitted to the electricity supplier.

6 Acknowledgment

This paper presents the scientific results of the project "Intelligent system for trading on wholesale electricity market" (SMARTRADE), co-financed by the European Regional Development Fund (ERDF), through the Competitiveness Operational Programme (COP) 2014-2020, priority axis 1 – Research, technological development and innovation (RD&I) to support economic competitiveness and business development, Action 1.1.4 - Attracting high-level personnel from abroad in order to enhance the RD capacity, contract ID P_37_418, no. 62/05.09.2016, beneficiary: The Bucharest University of Economic Studies.

References

[1] Palensky P., Dietrich D., *Demand Side Management: Demand Response, Intelligent Energy Systems, and Smart Loads*, IEEE Transactions on Industrial Informatics, Vol. 7, No: 3, Aug. 2011, pp. 381 – 388.

[2] Logenthiran T., Srinivasan D., Shun T.Z., *Demand Side Management in Smart Grid Using Heuristic Optimization*, IEEE Transactions on Smart Grıd, Vol. 3, No: 3, Sep. 2012, pp. 1244 - 1252.

[3] M. Shahidehpour, H. Yamin, and Z. Li, *Market Operations in Electric Power Systems*, New York: Wiley, 2002.
[4] L. Wu, M. Shahidehpour, T. Li,*Stochastic Security-Constrained Unit Commitment*, IEEE Transactions on Power Systems,vol. 22, no. 2, pp. 800-811, 2007
[5] J. Valenzuela, M. Mazumdar, *Monte Carlo computation of power generation production costs under operating constraints*, IEEE Trans. Power Syst., vol. 16, pp. 671-677, Nov. 2001.
[6] O.B. Tor, A.N. Guven, M. Shahidehpour, *Congestion-Driven Transmission Planning Considering the Impact of Generator Expansion*, IEEE Trans. on Power Systems, vol. 23 no. 2, pp. 781-789, May 2008.
[7]A. Bâra, S.V. Oprea, I. Șimonca (Botha), O.B.Tör, *Conceptual design and architecture of an informatics solution for smart trading on wholesale energy market in Romania*, Database Systems Journal vol. VII, no. 4/2016
[8] https://en.wikipedia.org/wiki/Pumped-storage_hydroelectricity#/media/File:Pumpspeicherkraftwerk _engl.png

Simona-Vasilica OPREA is an Assistant at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Bucharest University of Economic Studies of Bucharest. She has graduated the Polytechnic University in 2001, holds a Master Diploma in Infrastructure Management Program, Yokohama National University, Japan in 2007 and a PhD diploma from 2009. She is the author of over 30 articles, 3 of them ISI indexed. She coordinates a research project funded by NASRI. Domains of competence: databases, optimization, renewable power plants operation, smart grids, etc.

Osman Bülent TÖR received the B.S., M.S. and Ph.D. degrees from Middle East Technical University (METU), Turkey, in 1998, 2001 and 2008, respectively. He has almost 20 years working experience in the electrical power sector. At this moment, he is working as Managing Partner and Director for EPRA. Prior to this position he was leading the Power Systems Analysis and Planning Department of TÜBİTAK-Ankara, as a chief power system R&D engineer for 12 years, before he left TÜBİTAK and joined to EPRA in 2013. He has been lecturing as a part-time instructor at the METU - EE 578:Power System Planning, EE 573: Stability and Dynamics of Power Systems. He is the author of several journals and conference papers in the power system area.

A Modeling methodology for NoSQL Key-Value databases

Gerardo ROSSEL, Andrea MANNA
Universidad de Buenos Aires.
Facultad de Ciencias Exactas y Naturales.
Departamento de Computación. Buenos Aires, Argentina.
grossel@dc.uba.ar, amanna@dc.uba.ar

In recent years, there has been an increasing interest in the field of non-relational databases. However, far too little attention has been paid to design methodology. Key-value data stores are an important component of a class of non-relational technologies that are grouped under the name of NoSQL databases. The aim of this paper is to propose a design methodology for this type of database that allows overcoming the limitations of the traditional techniques. The proposed methodology leads to a clean design that also allows for better data management and consistency.
Keywords: *NoSQL, Key-Value Store, Conceptual modeling, NoSQL, Database design, Big Data*

1 Introduction

The need for analysis, processing, and storage of large amounts of data has led to what is now called Big Data. The rise of Big Data has had strong impact on data storage technology. The challenges in this regard include: the need to scale horizontally, have access to different data sources, data with no scheme or structure, etc. These demands, coupled with the need for global reach and permanent availability, gave ground to a family of databases, with no reference in the relational model, known as NoSQL or "Not Only SQL" (in some contexts also called NoSQL data stores).
There is a huge variety of NoSQL databases. They can be classified, among other things, according to the way they store and retrieve the information [1], [2]:

- Key-Value databases.
- Document databases.
- Column Families databases.
- Graph Databases.

Redis, one of the emblematic examples of key-value stores, is one of the most popular databases [3]. The development of conceptual modeling and general design methodology associated with the construction of NoSQL databases is at an early stage. There is previous work on development methodologies we can cite, like the Big Data Apache Cassandra methodology, proposed by Artem Chebotko [4].It uses the Entity Relationship Diagram as a conceptual model but it is oriented to a specific engine, Apache Cassandra. Thus, it is not generic and does not adapt to a design of key-value stores. Another proposal using a conceptual model for the design of NoSQL is described in [5]. It suggests the use of the various NoSQL databases common features to obtain a general methodology, in which an abstract data model called NOAM is used for conceptual data modeling. Such data model is intended to serve all types of NoSQL databases using a general notation.
Recently, an attempt to generate a universal modeling methodology adapted to both relational and non-relational database management systems was also presented, on the grounds of overcoming the constraints that the entity relationship model has, according to the author [6]
These efforts show that traditional methodologies and techniques of data modeling are insufficient for new generations of non-relational databases. It is necessary, then, to develop modeling techniques that adapt to these new ways of storing information. In this sense, this

paper will provide the tools to solve these limitations for key-value databases.
The rest of the paper is organized as follows: Section 2 outlines the fundamentals of the methodology proposed; Section 3 describes the main elements of the key-value databases design, which we call dataset and keeps certain correspondence with the entities of the conceptual model, to specify, in turn, the final adjustments that involve navigability between datasets and the addition of auxiliary datasets; Section 4 presents a simple design example and finally Section 5 presents conclusions and future work.

2 Methodology
The proposed design methodology has as its starting point the conceptual model, which can also be seen as a high-level description of data requirements. Conceptual modeling is usually performed using some form of entity-relationship diagram (DER) or conceptual class diagram in UML [7,8,9,10,11].

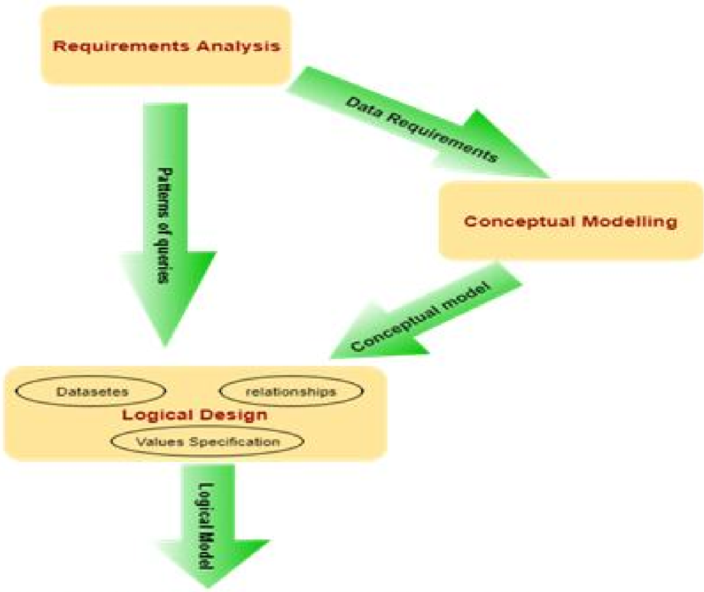

Fig. 1 Main Phases of Key Value Databases design

In traditional relational database design methodologies, conceptual modeling gave way to a logical design that was later transformed into a physical design.
It operates by transforming models from higher levels of abstraction to a model that maps directly into the structures of the database.
NoSQL databases, and in particular key-value storage systems, are generally referred to as *schema-less*, which seems to suggest that it is not necessary to make a model before the development starts. The fact that the structure of the data does not need to be defined has a priori many advantages for prototyping or exploratory development, but as data expands and the applications make use of them, the necessity to have them organized in some way arises.
In order to model and implement a key-value database correctly, it is necessary to take into account, firstly, the access patterns established in the requirements. In other words, the inputs to the logical and physical design are: the conceptual data model and the access patterns. It is assumed that during the phase of gathering requirements the feasibility of using this type of storage management system was analyzed.
From the conceptual model and access patterns, it is possible to specify the entities that are represented in the database by means the concept of datasets, as more precisely defined in section 3. Access to the values is performed using the keys. For every dataset it is necessary to realize a

correct design and keys selection. The keys design must be able to balance legibility and easy maintenance (to provide a logical structure), with access and storage efficiency. The keys also play an essential role in the implementation of distributed and scalable architectures, because they are used to distribute data across a set of servers.

Once the initial datasets have been established, the interrelations between them must be defined; indicating how the information obtained from a dataset can possibly enable access to the elements of another. In this context new datasets may arise whose role is simply to organize keys and favor navigability of the applications accessing the database. The diagram in Figure 1 illustrates the steps of the methodology proposed.

2 Datasets

Although this type of databases NoSQL is presented as a collection of key value pairs in such a way that access to the values is performed using a key, or part of it, it is also true that the data can be grouped into sets of data that represent some type of entity, relationship or event category in the real world.

These groups of data that can be inferred from the conceptual model are called datasets. All the elements (concrete elements) of a dataset share a common access mechanism, using a key with a uniform format. In addition, they present a similar scheme although there may be some differences between the elements of the same dataset. Namespaces, if any in the selected database engine, are not sufficient to differentiate the various types of datasets.

When in order to define a key, a data type is used that can adopt the same value for different datasets, a clash of keys could occur, i.e., there would be two equal keys referring to different values.

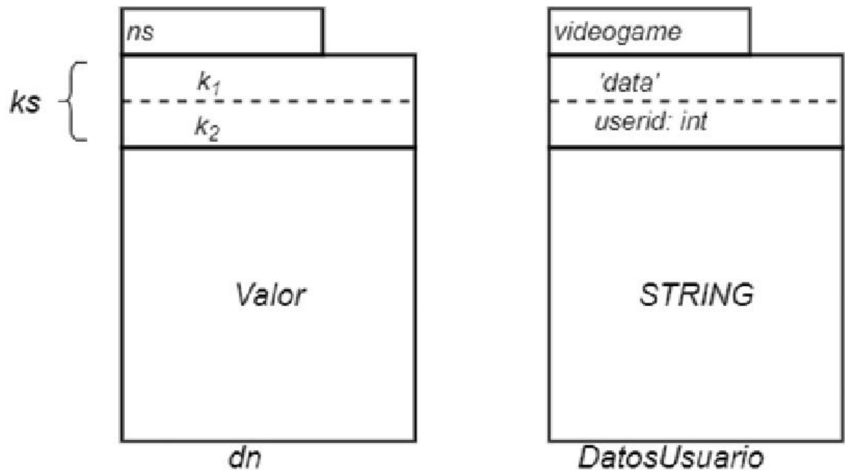

Fig. 2 Datasets Graphic Representation

For example, if you use a numeric key for customers and products, the value 2120 could be either the client 2120 or the product 2120. One way to avoid the clash of keys is to use a prefix that identifies the entity or context to which the key belongs. It is usual to use the name of the dataset as a prefix. This way a key "product:2020" would not conflict with another one "customer:2020". The prefixes that represent different entities allow for the identification of datasets.

The methodology in this article proposes that the logical design is represented by a set of interrelated datasets, defined as a 4-tuple, namely:

$dataset := \langle dn; ns; ks; v \rangle;$

where a dataset has a name dn, a namespace ns, an ordered list of elements that form the key ks, and a structure or a value type$v$.
The name "dn" allows to identify the dataset and must be unique, usually if it corresponds to an entity in the conceptual model same name is used. Namespace can be a string or a null value. In database engines that do not have support for namespaces its value becomes part of the key. The set of key parts ks helps usassemble complete key to access the value by concatenating them and using a separator between them. Each $k_i \epsilon k_s$ can take either a variable of some kindor constant value. The constant value is used, as previously explained, to avoidkey collision.
In our example of products and customers, it would be:

$$ks_{product} = \{k_1, k_2\}$$

where $k_1 = 'product'$ and $k_2 = idProduct: int$

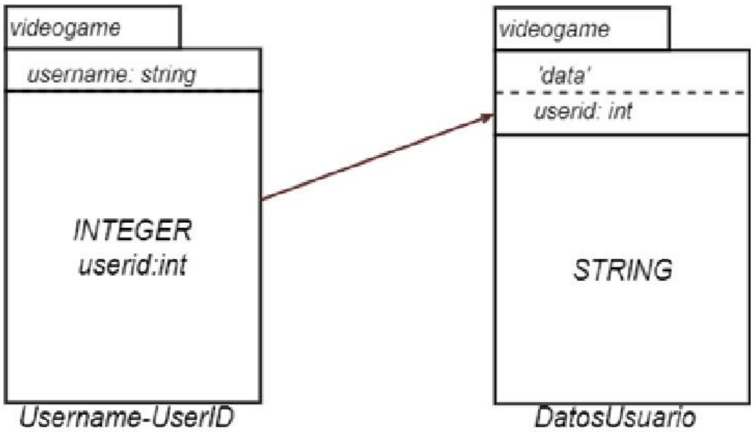

Fig. 3 Relationship between datasets

It is said that a concrete element e (a key-value pair) belongs to a *dataset* *(*$Ds := \langle dn; ns; ks; v\rangle$) if its key e_k is the ordered concatenation (with a uniform separator) of any of the possible values of each $k_i \epsilon k_s$ and its value (e_v) has the structure or the type of v. The value that takes the part of key k_i for a concrete element e_j is notatede_{jk_i} and its valuee_{jv}.
A navigation interrelationship between two datasets Ds_1 and Ds_2, indicates that an element of $e_1 \in Ds_1$ (a key-value pair or element) is related to an element $e_2 \in Ds_2$ and is defined as follows:

$$Ds_1 := \langle dn_1; ns_1; ks_1; v_1\rangle$$
$$Ds_2 := \langle dn_2; ns_2; ks_2; v_2\rangle$$
$$rel := \langle (Ds_1, Ds_2), lstVinc\rangle$$
$$lstVinc := (k_2 | (k_1, k_2))^*$$

where

$$k_1 \in ks_1 \; and \; k_2 \in ks_2$$

The ordered pair (Ds_1, Ds_2) indicates the direction of the navigability from Ds_1 to Ds_2 . The list of links $lstVinc$ defined the interrelationship, ie., how to obtain, from a concrete element $e_1 \in Ds_1$, the information to assemble the key to have access to a concrete element belonging to $e_2 \in Ds_2$. If there is only onekey in the list (which would correspond to Ds_2) then its value is obtained fromthe value of e_1. In the other case, the value takes the part corresponding to k_1 in e_1, whose value must be equal to the value taken by k_2 in element e_2, indicatingthat the concrete elements in each of these parts of the key have the same value.
The interrelationship is completely defined by all the elements of the list $lstVinc$
The datasets and their relationships can be

represented in a diagram to visualize them and the navigation between them immediately. Figure 2 shows the graphic representation of a dataset, as well as a concrete example.
The interrelations that allow from the value or a part of the key of a concrete element in a dataset access to an element in another dataset is represented by directed arrows departing from the value or key part of the first dataset to the key part *k* corresponding to the set *ks* of the second one. Figure 3 shows a relationship between two datasets, the one labeled *UserName-UserID* has a numerical identifier or *userid* by each user name (used as key) that can be used in other places to access the information that corresponds to that user. The*userid* can be used, for example, to access user data such as email, address, name, and surname, etc. stored in another dataset.
This relationship example is a verycommon pattern in key-value databasesthat allows such things as renaming a user with minimal impact or easily generating reverse indexes

4Application Example

In this section, a simple partial example is presented to illustrate how a diagram is done. The example is to design a key-value database that stores the information of users and online game matches. The input to dataset design is the conceptual model, query patterns and requirements. To keep the example simple only a small part of the problem domain is shown. For the requests, it is known that every user can save a game at any time. Upon entering the game, the user can choose to continue from any state where the game was saved.
The status of the items is determined by a location in the map of the game, a group of objects and other data as the level of energy, remaining lives, etc.
From the conceptual model, the following entities and relationships are derived: users, starting status and friends (one user can be a friend of another). The established query patterns are: view the list of saved games, retrieve a saved game and view a user's friends.

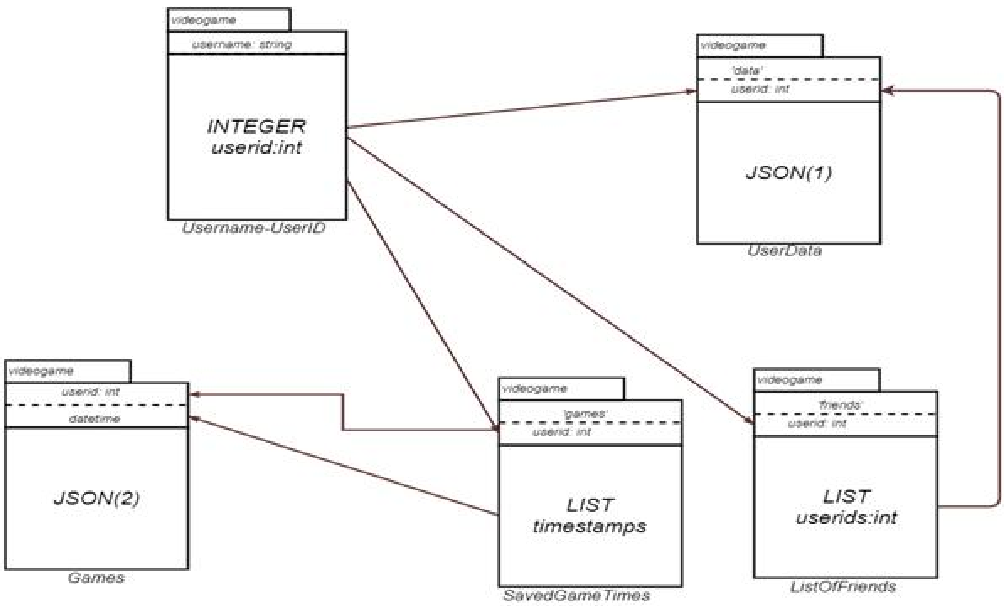

Fig. 4 Logical Design Example

The datasets arising from this simplified description represent items and users (entities of the conceptual model). One of the ways to solve the interrelations that exist in the conceptual model is to use a dataset that allows one entity to navigate to

another one. In the simplified example presented, a dataset is used to store a list of timestamps that would show the moments in which a user saved their games. To identify the user the pattern outlined in the previous section can be used, which links a user name with an identifier. Figure 4 shows the complete diagram. Labeling values as type JSON with an associated number would specify
the structure that this JSON would take in the format *JSON Schema.*

5 Conclusions

This article has presented an original methodology to support the design of NoSQL Key-Value databases. The input to the logical design incorporates query patterns along with the conceptual model. We have explained how to represent the conceptual entities and design through the concept of dataset. In addition, a new type of diagram was specified, which allows the specification of datasets and interrelations established during the design phase.
These are some highlights of the proposed methodology:

- This is, to the best of our knowledge, the first work that presents a method logy supported by specific notation to develop NoSQL databases based on key-value.
- The diagrams presented favors communication and understanding of the design decisions made during the development

The prospects for this project involve the improvement of this work, based on the experience of its use in various development projects and the production of a series of transformation patterns from the conceptual model to the logical design.

6Acknowledgment

We want to thank María Jose Constanzo of University of Buenos Aires for their assistance in the final writing of this article.

References

[1] Adam Flowler,"The State of NoSQL", 1st edition, 2016

[2] Pramod J. Sadalage, Martin Fowler,"NoSQL Distilled: A Brief Guide to the Emerging World of Polyglot Persistence", *Addison-Wesley Professional*,1st edition, 2012

[3] "DB-Engines Ranking"http://db-engines.com/en/ranking (consulted in June 2017)

[4] ArtemChebotko, Andrey Kashlev, Shiyong Lu, "A Big Data Modeling Methodology for Apache Cassandra", *IEEE International Congress on Big Data (BigData'15)*, pp. 238-245, New York, USA, 2015.

[5] Francesca Bugiotti, Luca Cabibbo, Paolo Atzeni, Riccardo Torlone. "Database Design for NoSQL Systems". *International Conference on Conceptual Modeling*,pp.223 - 231 Atlanta, USA, Oct 2014.

[6] Ted Hills,"NoSQL and SQL Data Modeling",*Basking Ridge, NJ: Technics Publications*, 2016

[7] Peter P. S. Chen,"The entity-relationship model: toward a unified view of data",*Proceedings of the 1st International Conference on Very Large Data Bases,* ACM, New York, NY, USA, 1975.

[8] Carlo Batini, Stefano Ceri, and Shamkant B. Navathe, "Conceptual Database Design: An Entity-Relationship Approach", *Benjamin-Cummings Publ. Co., Inc.,* Redwood City, CA, USA, 1991

[9] François Lagarde, Huáscar Espinoza, François Terrier, and Sébastien Gérard, "Improving uml profile design practices by leveraging conceptual domain models.", *Proceedings of the twenty-second IEEE/ACM international conference on Automated software engineering (ASE '07)*, pp. 445-448 ACM, New York, NY, USA, 2007

[10] Antoni Olivé. "Conceptual Modeling of Information Systems.",*Springer-*

Verlag New York, Inc., Secaucus, NJ, USA, 2007

[11] Toby J. Teorey, Dongqing Yang, and James P. Fry,"A logical design methodology for relational databases using the extended entity-relationship model",*ACM Comput. Surv*. 18, 2, June 1986

Gerardo ROSSEL graduated as Ms. Sc. in Computer Science from the Faculty of Exact and Natural Sciences of the University of Buenos Aires. He has a Doctor's degree from the National University of Tres de Febrero. At present, he is an assistant lecturer at Computer Department of FCEyN UBA. He has more than 20 years of experience in software industry and is Chief Scientist of UpperSoftsoftware company. Her specific area of competences is in Software Engineer, Databases, NoSQL, Intelligent Agents and Multiagent Systems, Patterns and Software Architectures, Epistemology and History of Computer Science. He is co-author of book "*Algoritmos,Objetos y Estructuras de Datos"*. He has published several papers in national and international conferences and journals. He was a member of the International Program Committeeof several international conferences.

Andrea MANNA, graduated from the Faculty of Exact and Natural Sciences of the University of Buenos Aires in 2000. She got the title of Ms. Sc. of Computer Science. At present, she is Head of practical work in the Faculty of Exact and Natural Sciences of the University of Buenos Aires. She has been working in the software industry since 1995. She is Chief Software Architect of UpperSoftSofware Company and work in software development for more than twenty years. She is co-author of book "*Algoritmos,Objetos y Estructuras de Datos"*.

Databases in Cloud – Solutions for Developing Renewable Energy Informatics Systems

Adela BÂRA[1], Iuliana ȘIMONCA (BOTHA)[2], Anda BELCIU (VELICANU)[3]
[1,2,3]The Bucharest University of Economic Studies
{bara.adela, iuliana.botha, anda.velicanu}@ie.ase.ro

The paper presents the data model of a decision support prototype developed for generation monitoring, forecasting and advanced analysis in the renewable energy filed. The solutions considered for developing this system include databases in cloud, XML integration, spatial data representation and multidimensional modeling. This material shows the advantages of Cloud databases and spatial data representation and their implementation in Oracle Database 12 c. Also, it contains a data integration part and a multidimensional analysis. The presentation of output data is made using dashboards.

Keywords:*Renewable Energy, Data Model, Databases in Cloud, Spatial Data, Data Warehouse*

1 Introduction

The main objective of the research project Intelligent System for prediction, analysis and monitoring of performance indicators of technological and business processes in the field of renewable energies (SIPAMER) is to develop a prototype of a decision support system for energy producers form renewable resources like wind and solar sources. The prototype's architecture, more detailed in [1] consists of three layers as shown in **Fig. 1**.:

- data layer that collects data from measuring devices installed in the wind/photovoltaic power plants and from SCADA/EMS systems used for monitoring the energy power produced by each turbine or photovoltaic panel;
- models layer containing the data mining algorithms with artificial neural networks (ANN) for energy production forecast and also with multidimensional OLAP models for decision support;
- interface layer with business intelligence capabilities that enables advanced and interactive analysis of key performance indicators through dashboards and dynamic reporting tools.

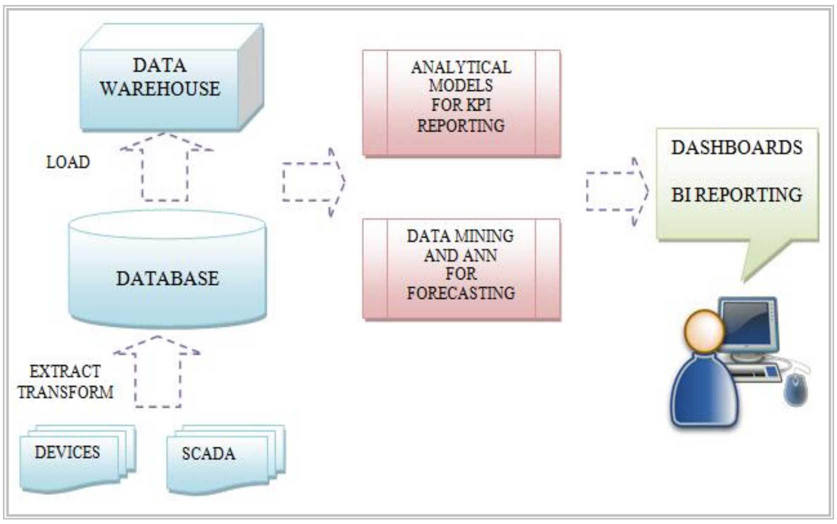

Fig. 1. SIPAMER's architecture

The producers' requirements in terms of real time analysis and accurate forecasting regarding energy production are implemented in two modules accessible through a cloud computing platform. We set up a database instance dedicated to each producer with Business Intelligence reporting tools accessible as a service via secured connection. The database instance is Oracle Database 12c that allows advanced XML and JSON integration, spatial data representation, multidimensional and OLAP analytics.

2 Cloud Databases

No matter what the business is about and what software it uses, most likely it will be faced with dealing with the concepts of Big Data, Cloud, Smart and Business Rules [2], [3].

Cloud computing is a paradigm that allows access through network to a pool of resources dynamically configurable, which are available to everyone as services. These resources can be quickly found and can be provided with minimal effort through a service provider interaction. This model's first operating principle refers to resource availability.

This innovative paradigm and also the large volumes of different types of data, which involve new data storage requirements, all have led to the cloud databases.

The database systems have as foundation the data level (which can be received as a Big Data source), represented by the database. Cloud computing offers an efficient manner of processing and managing large volumes of data, since the big data philosophy involves gathering and processing massive data sets which are so large that conventional database systems and software tools have failed to manage them.

Oracle 12c is designed especially for cloud environment. It offers a Cloud solution, which brings new improved features like providing the database as a service in Cloud, optimizations, integration and analysis of Big Data, security and so on [4]..

Cloud databases have a number of advantages, among which saving the storing space, by using the cloud storing, is the most important. Of course, this also comes with some protection risks (integrity and security) depending on a third party's measures of security. However, if the private company is not specialized in providing database protection, probably the cloud providers do it better then than average IT companies, having a background on the subject. Another impediment can be the money spent on storing data in a Cloud database. However, they became more and more price accessible and will make it worth, by providing specialized hardware, like database machines. This is how personal hardware resources aren't a problem anymore. The only problem remains the internet dependency, which necessitates a good speed of traffic.

3 Data integration

Data layer integrates data from SCADA/EMS connectors and flat files from measuring devices that transmit temperature, atmospheric pressure, wind speed and direction, humidity and solar radiation. The integration process involves routines that extract and transform flat files into XML pattern files and import data into relational tables represented as XSD schemas. For example, the measuring data for photovoltaic panels are modeled into DATA_METEO_PV XSD schema as shown in **Fig. 2.**

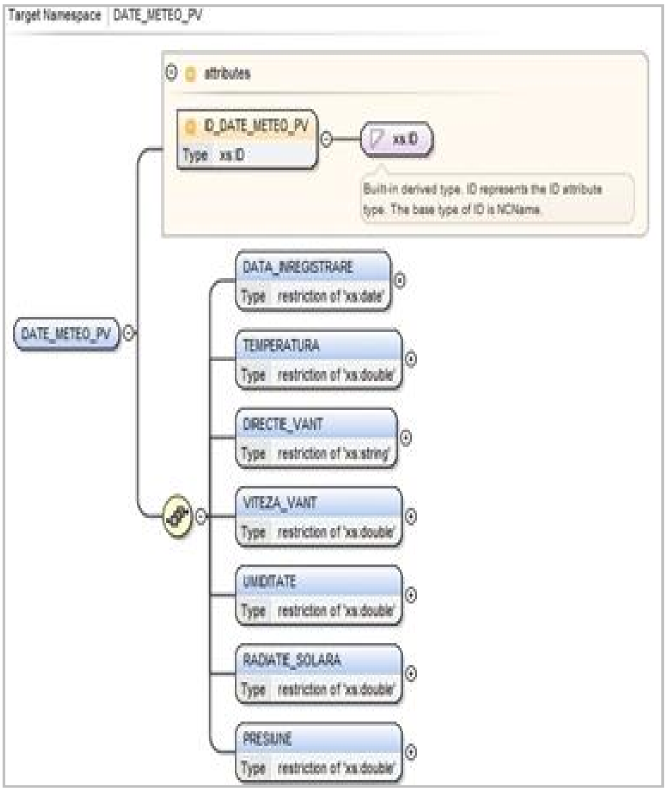

Fig. 2. XSD schema for photovoltaic panels' measuring data

The advantage of the model is that the import process is fast and flexible [5]; any change can be easily captured inside the XML patterns.

The XSD Schema for photovoltaic panels' measuring data can be implemented using specific sintax, as follows:

```
<?xml version="1.0" encoding="utf-8"?>
<xs:schema
xmlns:xs="http://www.w3.org/2001/XMLSc
hema" xmlns="date_meteo_pv"
targetNamespace="date_meteo_pv">
<xs:element name=" date_meteo_pv ">
<xs:complexType>
<xs:sequence>
<xs:element name="data_inregistrare">
<xs:simpleType>
<xs:restriction base="xs:date">
</xs:restriction>
</xs:simpleType>
</xs:element>
<xs:element name="temperatura">
<xs:simpleType>
<xs:restriction base="xs:double">
</xs:restriction>
</xs:simpleType>
</xs:element>
. . . . . . . . . . .
<xs:element name="presiune">
<xs:simpleType>
<xs:restriction base="xs:double">
</xs:restriction>
</xs:simpleType>
</xs:element>
</xs:sequence>
<xs:attribute name="id_pv"
type="xs:integer" use="required"/>
</xs:complexType>
</xs:element>
</xs:schema>
```

Further, we used some mapping algorithms, in order to transform XSD Schema into database objects (tables), so that uploaded data can be validated in compliance with the schema.

The database schema for each producer's instance contains relational tables for photovoltaic panels, wind turbines, meteorological recordings, production, operation and maintenance data and transactions on energy market. We also set up a dedicated instance for public actors like energy market regulators (ANRE and OPCOM) or national TSO (Transelectrica SA). The instance provide a process that imports the mandatory reports of energy producers (**Fig. 3.**) into a relational schema, aggregate them and make data available for multidimensional analysis.

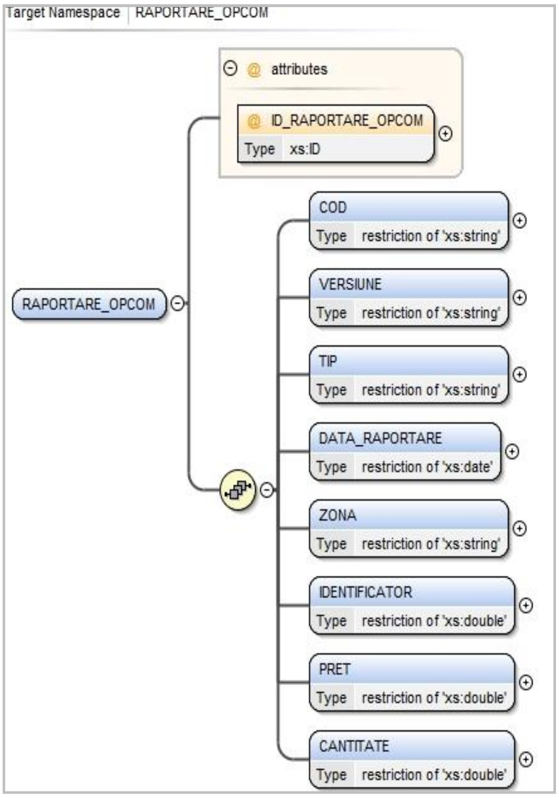

Fig. 3. XSD schema for OPCOM reporting

The XSD schema is implemented in Oracle Database 12c environment and enriched with spatial data representation in order to develop interactive charts and reports.

4 Spatial data representation

GIS (Geographic Information Systems) are based on spatial databases. Geographic Information Systems are capable of storing, manipulating, analyzing and displaying geographically referenced date towards a coordinate system (standard or user-defined), according to [6]. The most important Web GIS applications are described in [7] and the spatial analysis is the subject of [8] book.

Some domains (like transportation, energy, tourism, meteorology, etc.) need spatial representation of information on maps, by default. Because of the growing capacity of storing data in databases, in cloud and in cloud databases, we are allowed to store as much data as we need it. For example, storing all the coordinates that compound the perimeter of a lake was a problem of space a few time ago, but not anymore. Now, the only concern is how to make the analysis of spatial data go faster and return interactive results.

This is the problem that Oracle Database 12c solves quite well by grouping the spatial elements in a set of technologies that offers support starting from storing data, to analyzing and representing it. Oracle Location Technologies gather together Oracle Database Locator feature, Oracle Spatial and Graph Option and Oracle Fusin Middleware Map Viewer.

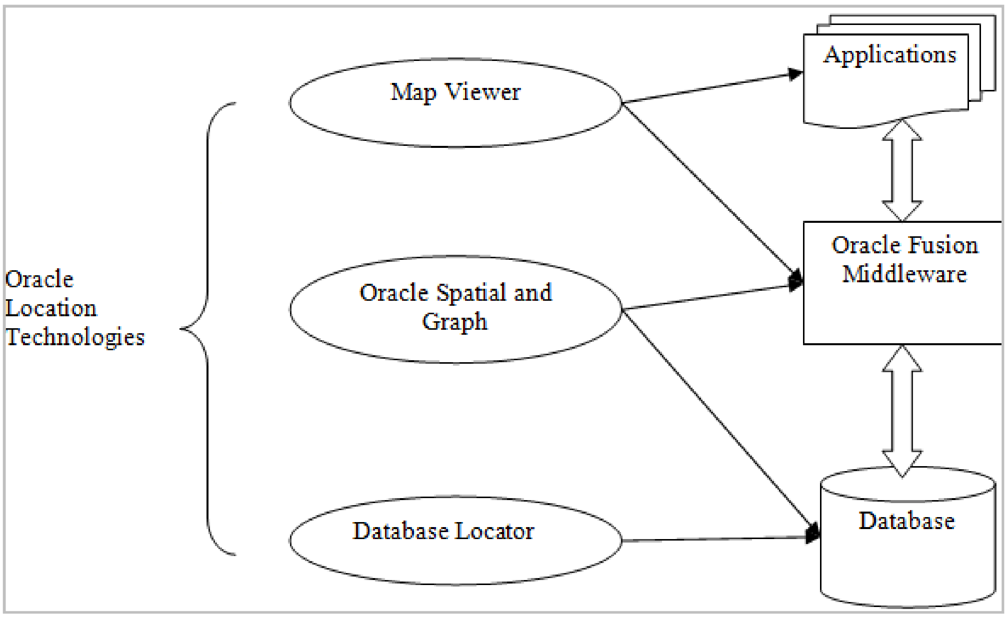

Fig. 4. Oracle Location Technologies

Fig. 4.points out the three components of Oracle Location Technologies and their interaction with the layers of an Oracle System. Database Locator allows storing spatial data in a specialized data type called SDO_GEOMETRY. Points, lines and polygons can be stored and queried using spatial operations.

Oracle Spatial and Graph (called Oracle Spatial until now) adds more spatial capabilities to Oracle Locator by using spatial and graph analysis. This type of analysis comes with the Database engine but it is used in applications through the middle layer of database system architecture. Also, the spatial queries based on spatial joins, touches, overlaps and other spatial relations, are run fifty times faster [6].

The Map Viewer develops web-based maps using data stored in Oracle Database with Spatial features. It is included in many Oracle products like Business Intelligence, SOA Suite, E-business Suite, JDeveloper and so on. Although Map Viewer it is considered to be part of the Oracle Fusion Middleware, by simplifying the complexity of an application, it can be said it provides the final application interface.

When working with SDO_GEOMETRY type in Oracle, the following steps must be followed:

- creating the spatial tables that use the spatial data
- inserting spatial data by using the SDO_GEOMETRY function, which has the following syntax:

```
  SDO_GEOMETRY(
    2003,
    NULL,
    NULL,
SDO_ELEM_INFO_ARRAY(1,sdo_etype,sdo_in
terpretation),
SDO_ORDINATE_ARRAY(x1,y1,x2,y2,…)  )
```

The combination of parameters sdo_etype, sdo_interpretation gives the type of stored object. For example, if the parameters have the values 1003, 3, they will define a rectangle. In bi-dimensional coordinates, the function SDO_ORDINATE_ARRAY() will get (xi,yi) pairs of parameters, depending on the type that was established through SDO_ELEM_INFO_ARRAY. For example, for a rectangle there will be given two pairs of coordinates.

- inserting a new record in the standard metadata view USER_SDO_GEOM_METADATA, which has the effect of defining the coordinate system, whether is a standard or a user-defined one;
- creating the spatial indexes is necessary for all spatial columns in all spatial tables. Spatial indexes are R-tree indexes and are used in requests that use spatial criteria like areas, perimeters, distance, union of spatial objects, etc.

The most common spatial operations are: create, manipulation (insert, update, delete), queries, map operations (overlapping layers, reducing the number of coordinates that define a spatial object, etc.), conversions. The spatial queries can be based on operations for determining the spatial relations or on spatial analysis operations (see [10]).

In Oracle, the spatial operators are implemented using certain functions like SDO_AREA, SDO_LENGTH, SDO_UNION, SDO_INTERSECTION, SDO_DIFFENECE, SDO_CENTROID, etc.

5 Multidimensional analysis

To develop multidimensional analysis model we designed snowflake schema and implemented in Oracle Warehouse Builder. The solution adopted for the prototype is based on cube based multidimensional model (MOLAP) implemented in the cloud computing platform. Thus, on the instance dedicated to energy producers we developed cubes for production analysis, forecasting and financial simulation and on the instance dedicated to public authorities we developed cubes for production forecasting analysis on regional/national level.

The steps for developing included the data cleansing, extract, transform and load process (ETL), mappings, validations and multidimensional objects generation: dimensions (producers, locations, regions, time, power plant, substations) and cubes (contracts, forecasting, production, weather, energy market transactions).

Data cleansing was mostly realized through data rules derived from the process of data profiling. In order to perform data profiling we need to follow some steps (detailed in [12]), which involve creating data profile objects, developing and configuring the profiles, loading the configuration parameters (such as: domains, functional dependencies, patterns etc.) and then executing the data profiling process.

We started by checking the data sources used, in order to establish common formats.

By following the steps mentioned above, we present an example of data profiling made in Oracle Warehouse Builder.

From the large SIPAMER's database schema we have selected the Investors table (see the structure in the Fig.5).

We can note the following inconsistencies in the data stored:

- tara: multiple versions of storing the same data;
- banca_investitor: some values are incorrect;
- telefon: multiple versions of storing the same data.

After running the data profile process, they will be displayed detailed statistics about the data. There will be also detected the domain values for Tara and Telefon columns. Based on these values, OWB will derive data rules which will be applied as data integrity restrictions in database tables.

The following PL/SQL function it is used to achieve country-level corrections:

```
begin
if upper (tara) in ('RO','ROU') then
return 'ROMANIA';
      elsif upper (tara) = 'SZ' then
return 'ELVETIA';
      elsif upper (tara) = 'UK' then
return 'ANGLIA';
      else                    return
upper(NVL(tara,'NECUNOSCUT'));
end if;
end;
```

Also, OWB will detect some incorrect values for the investors' bank. We need to select the cleanse strategy for the corrections, so we will choose to correct the data, according with the identified domain values.

We applied the profiling process for other tables such as bids, transactions, and beneficiary and also for the tables that gather data from meteorological devices. After the cleansing process we build in OWB the data warehouse schema (**Fig. 5.**) that is a snowflake schema with joins between dimensions and types of dimensions.

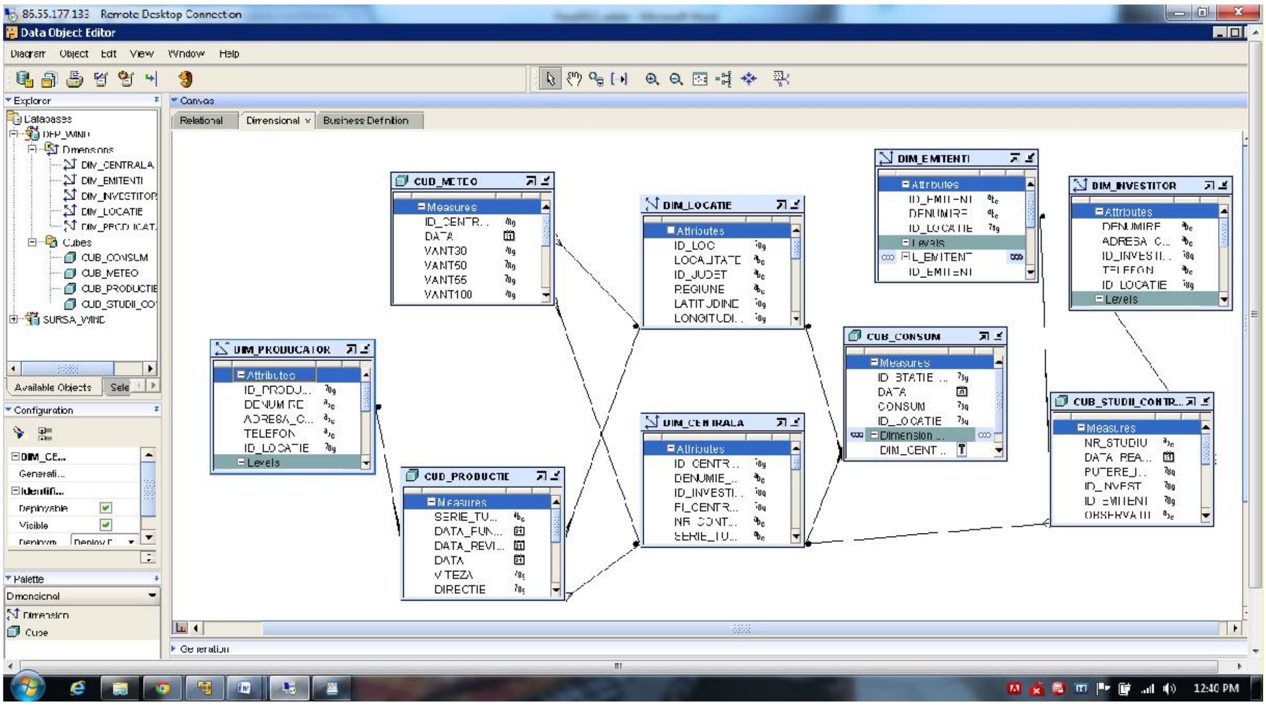

Fig. 5. Snowflake DW schema

In order to develop dashboards for multidimensional analysis, we build the following models in Oracle business Intelligence Suite: data model that contains the data locations and mappings, logical model based on the OWB data warehouse schema that contains dimensions, levels, facts and measures and presentation model that contains the attributes displayed in the dashboards and analytical reports.

6 Displaying data using dashboards

For each actor (energy producers and public authorities) we developed separate dashboards available through cloud services that offer a customized set of analytical reports regarding the production activities, forecasting for different time intervals (hours to 3-7 days), key performance indicators for energy market transactions (green certificates, intra-day, next day and balancing energy markets) as detailed in [7].

For public authorities the dashboard include a section dedicated to correlations between wind or photovoltaic power plants generation in the same region/area and aggregate regional or generation type forecasting.

Using the dashboards the decision factors have access to analytical reports from the system, such as:

- Reports for the contracts in progress in terms of installed capacity in wind or photovoltaic power plants by regions and by coupling stations;
- Reports to analyze production on different periods and in different regions;
- Reports to analyze consumption in different time periods and different points of measurement;
- Reports to analyze the evolution of weather conditions and energy predictions.

Oracle BI Dashboard integrates all reports made in a dashboard accessible by both computers and laptops and mobile devices like PDA or mobile phones. The dashboard will have the following sections: Home, Contracts, Consumption analysis, and Production analysis, Predictions (**Fig. 6.** and **Fig. 7.**).

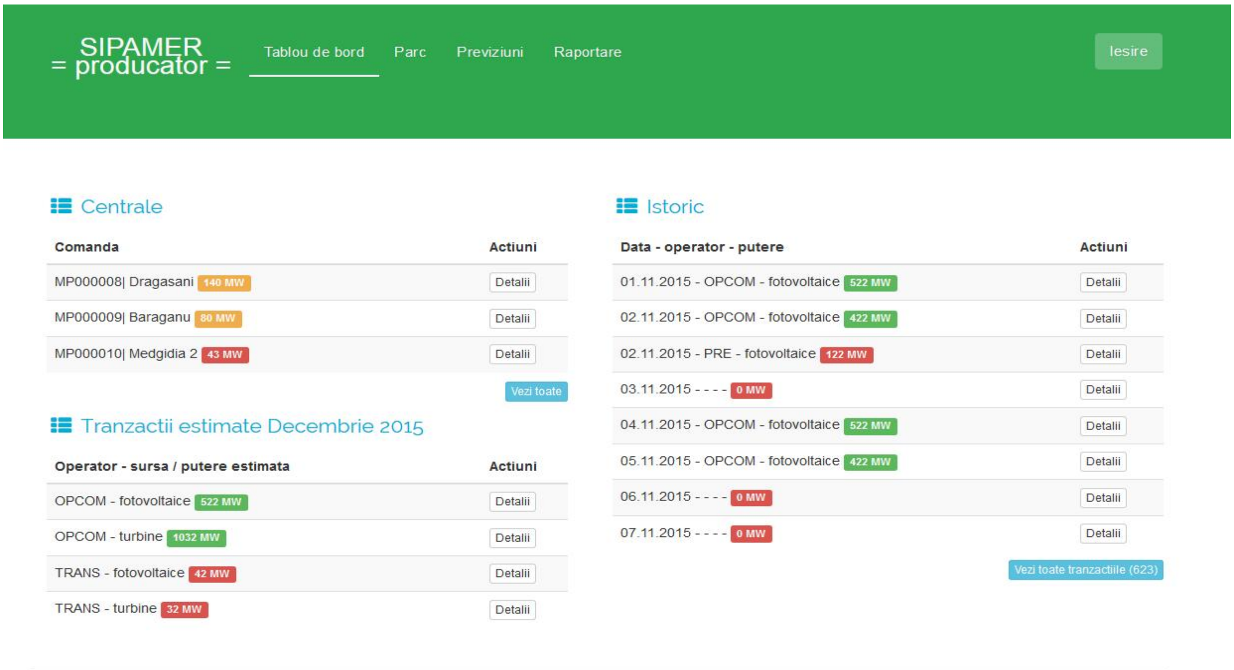

Fig. 6. Dashboard's first page

This method of presentation allows easy navigation and analyzing reports in a centralized manner, and also offers the opportunity to reconfigure reports (using OBI Answers) by modifying parameters or presentation style.

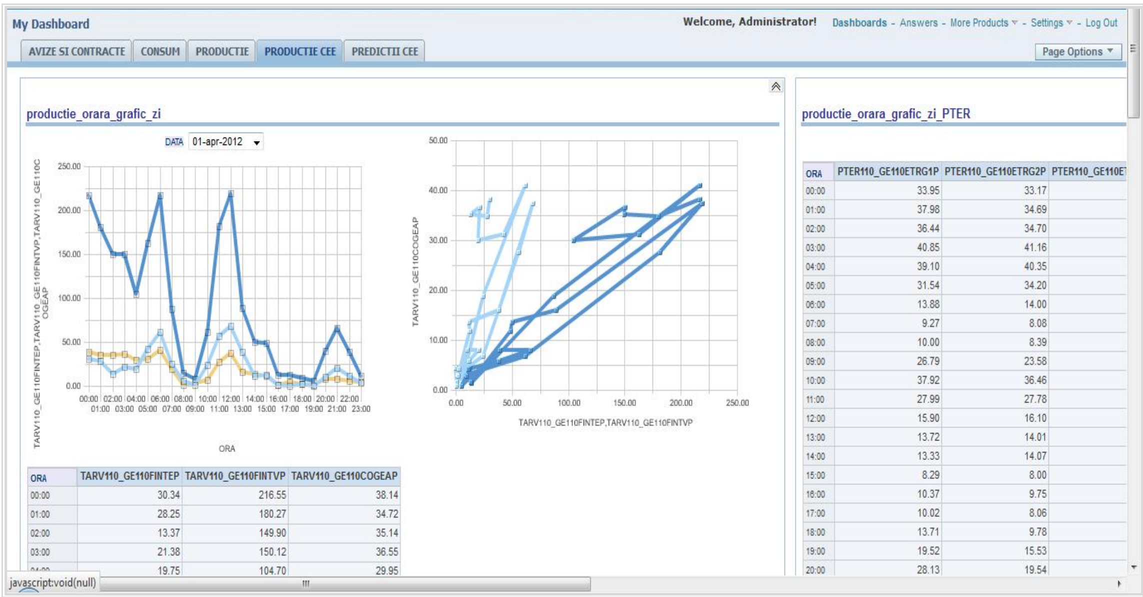

Fig. 7. Reports for analyzing the energy production, that are integrated into the dashboard

Fig. 7.presents the analysis section of the renewable production through interactive reports, charts that shows the degree of simultaneity in operation and aggregated reports for indicators regarding the production days / months and regions.

7 Conclusions

The SIPAMER prototype is a dedicated informatics solution for decision support in wind or photovoltaic power plants management both for producers and for public authorities. It is built on modules that include generation monitoring, forecasting and prediction the production based on meteorological factors and financial simulations regarding the energy market transactions. The cloud computing platform allows us to develop the prototype without infrastructure constrains, instantiating only the Oracle Database 12c for each actor and setting up the developing environment. The dashboards and monitoring tools are available to users through a portal based cloud service.

8Acknowledgment

This paper presents results of the research project: Intelligent System for prediction, analysis and monitoring of performance indicators of technological and business processes in the field of renewable energies (SIPAMER), research project, PNII –PCCA 2013, code 0996, no. 49/2014 funded by NASR.

References

[1] A. Bâra, S.V. Oprea, I. Lungu, G. Căruţaşu, C. P. Botezatu - Financial simulator based on wind power plants forecast, World Renewable Energy Congress 14, WREC XIV, University Politehnica of Bucharest, Bucharest – Romania, June 8 – 12, 2015, accepted for publishing in Open Access Journal of Physics: Conference Series (JPCS), 1742-6596

[2] S.V. Oprea, I. Lungu - Informatics Solutions for Smart Metering Systems Integration, RevistaInformaticăEconomică, vol. 19, nr. 4/2015 http://revistaie.ase.ro/content/76/03%20-%20Oprea,%20Lungu.pdf

[3] Alexandra Maria IoanaFlorea, Ana-Ramona Bologa, Vlad Diaconița, RazvanBologa - Streamlining Business Processes in Academia by Building and Manipulating a Business Rules Repository, The 14th International Conference on Informatics in Economy (IE 2015), ISSN 2284-7472, ISSN-L 2247-1480

[4] Oracle White Paper - Plug into the Cloud with Oracle Database 12c, October 2015, http://www.oracle.com/technetwork/database/plug-into-cloud-wp-12c-1896100.pdf

[5] H. Huiping Cao, Y. Qi, S. Candan, M.L. Sapino - XML Data Integration: Schema Extraction and Mapping. Advanced Applications and Structures in XML Processing: Label Streams, Semantics Utilization and Data Query Technologies, IGI Global, 2010

[6] Michael Brueckner, OrasaTetiwat, Use of Geographical Information Systems for Thailand, E-leader Bangkok, 2008.

[7] Pinde Fu, Jiulin Sun, Web GIS: Principles and Applications, Esri Press, 2010, ISBN:158948245X 9781589482456.

[8] Peter Rogerson, S Fotheringham, Spatial Analysis And GIS (Technical Issues in Geographic Information Systems), 296 pg., 2007, Taylor&Francis, ISBN-10: 0748401040.

[9] Oracle White Paper - Oracle Database 12c: An Introduction to Oracle's Location Technologies, September 2014, http://download.oracle.com/otndocs/products/spatial/pdf/12c/oraspatialandgraph_12c_wp_intro_to_location_technologies.pdf

[10] Belciu, A.: Baze de date spatiale in arhitecturaorientatapeservicii, ASE Publisher, Bucharest (2014)

[11] I. Lungu, G. Căruţaşu, A. Pîrjan, S.V. Oprea, A. Bâra - A two-step forecasting solution and upscaling technique for small size wind farms located on hilly areas in Romania, Studies in Informatics and Control Journal, vol 25/issue 1, 2016, ISSN 1220-1766

[12] Oracle Warehouse Builder Data Modeling, ETL, and Data Quality Guide 11g Release 2 (11.2), http://docs.oracle.com/cd/E11882_01/owb.112/e10935/data_profiling.htm

Adela BÂRA (b. October 11, 1978) is Professor at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Economic Cybernetics in 2002, holds a PhD diploma in Economics from 2007. She is the author of 9 books in the domain of economic informatics, over 40 published scientific papers and articles (among which over 20 articles are indexed in international databases, ISI proceedings, SCOPUS and 10 of them are ISI indexed). She participated as team member in 5 research projects and has gained as project manager 4 research grants, financed from national research programs. Domains of competence: Database systems, Data warehouses, OLAP and Business Intelligence, Executive Information Systems, Decision Support Systems, Data Mining.

Iuliana ȘIMONCA (BOTHA) is Associate Professor at the Economic Informatics Department at the Faculty of Cybernetics, Statistics and Economic Informatics from the Academy of Economic Studies of Bucharest. She has graduated the Faculty of Cybernetics, Statistics and Economic Informatics in 2006, the Databases for Business Support master program organized by the Academy of Economic Studies of Bucharest in 2008 and she holds a PhD diploma in Economic Informatics from 2012. She is author/co-author of 9 books and 44 papers published in journals or conference proceedings (4 papers published in ISI indexed journals, 14 included in international databases and 9 published in ISI proceedings). She participated as team member in 8 research projects (among which two international research programs). Her scientific fields of interest include: Database Systems, Data Warehouses, Business Intelligence, Design of Economic Information Systems.

Anda BELCIU (VELICANU) has graduated the Faculty of Economic Cybernetics, Statistics and Informatics of the Bucharest University of Economic Studies, in 2008. She has a PhD in Economic Informatics and since October 2012 she is a Lecturer. She teaches Database, Database Management Systems and Software Packages seminars and courses at the Economic Cybernetics, Statistics and Informatics Faculty. Her scientific fields of interest and expertise include database systems, e-business, elearning, spatial databases.

Analysis of value added services on GDP Growth Rate using Data Mining Techniques

Douglas KUNDA[1], Sipiwe CHIHANA[1]
[1]Department of Computer Science, School of Science Engineering and Technology, Mulungushi University, PO box 80415 Kabwe, Zambia
dkunda@mu.edu.zm, sipiwechihana09@gmail.com

The growth of Information Technology has spawned large amount of databases and huge data in numerous areas. The research in databases and information technology has given rise to an approach to store and manipulate this data for further decision making. In this paper certain data mining techniques were adopted to analyze the data that shows relevance with desired attributes. Regression technique was adopted to help us find out the influence of Agriculture, Service and Manufacturing on the performance of gross domestic product (GDP). Trend and time series technique was applied to the data to help us find out what trend of GDP with respect to service, agriculture and manufacturing sector for the past decade has been. Finally Correlation was also used to help us analyze the relationship among the variables (service, agriculture and manufacturing sector). From the three techniques analyzed, service value added variable was the most prominent variable which showed the strong influence on GDP growth rate.

Keywords: *GDP, Regression, Time-series/trends analysis, Correlation, Data mining, Predictions*

1 Introduction

The development of Information Technology has produced large amounts of databases and huge data in numerous areas. The research in databases and information technology has given rise to an approach that stores and manipulate this data for further decision making. Data mining began its life in specialist applications such as geological research and meteorological research, more recently it has been applied in a number of areas of industry and commerce [1]. Data mining is a process of extraction of useful information and patterns from huge data, it is also called as knowledge discovery process, knowledge mining from data, knowledge extraction or data pattern analysis [2]. [3]Adds that Data mining generally refers to the process of extracting interesting hidden information from available chunks of data, which could otherwise be manually impossible. Despite data mining being relatively a new technology that has not completely matured, there are a number of areas that are already using it on a regular basis. Some of these organizations include retail stores, hospitals, banks, insurance companies. Many of these organizations are combining data mining with such things as statistics, pattern recognition, and other important tools [2]. Data mining in economics is an emerging field of high importance for providing prognosis and a deeper understanding of economic data. Researchers are using data mining techniques in the analysis and prediction of several economic indicators such as GDP. [3]) points out that GDP is one of the measures of national income and output for a given country's economy at a given period of time and adds that the definition of GDP is based on the total market value of all final goods and services produced within the country in a given period of time (normally one year). GDP is also an important statistic that indicates whether an economy is expanding or contracting. To know whether an economy is expanding or not and what sectors contribute the most to annual GDP growth and to find out how GDP has been performing in a given period of time can be achieved by using data mining techniques. Additionally Successful data mining is based on various investigations of the data using different methods, parameters, and data to find most meaningful relations [4]. [2]Various algorithms and techniques like Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbor method etc., are used for knowledge discovery from databases. Data mining techniques such as Trend analysis, application of forecast in terms of time series model is found widely in economic development [5], Regression and Correlation helps in finding the patterns, relationships among different economic sectors on how they influence GDP growth to decide upon the future trends on how GDP is likely to perform.

Based on the above information this paper seeks to discuss the analysis that was conducted using data mining techniques on GDP with a view to answer some data mining questions that were proposed to gain a deeper understanding of the dataset, below are the questions that guided the authors on what techniques to use;

1. What is the influence of Agriculture, Service and Manufacturing on the performance of gross domestic product (GDP)?
2. What has been the trend of gross domestic product with respect to service, agriculture and manufacturing sector for the past decade?
3. What is the relationship among the variables service, agriculture and manufacturing sector with respect to GDP as a controlling variable?

1.1. The Significance of GDP

Economic growth of a country is measured by the Gross Domestic Product (GDP) [6]. Gross domestic product (GDP) is the backbone of Zambia's economy and plays a significant role in that it is our economic indicator that shows how the economy performs with respect to many sectors and it also shows whether as a country we are growing or not. According to ([7]) GDP is one of the measures of national income and output for a given country's economy at a given period of time. [8] Adds that the definition of GDP is based on the total market value of all final goods and services produced within the country in a given period of time (normally one year). In a related literature [9] adds that GDP is the total market value of all final goods and services produced in a country in a given year, equal to total consumption, investment and government spending, plus the value of exports, minus the value of imports.GDP is commonly used as an indicator of the economic health of a country, as well as a gauge of a country's standard of living. Since the mode of measuring GDP is uniform from country to country, GDP can be used to compare the productivity of various countries with a high degree of accuracy. Adjusting for inflation from year to year allows for the seamless comparison of current GDP measurements with measurements from previous years or quarters. In this way, a nation's GDP from any period can be measured as a percentage relative to previous periods. [10] GDP is total national income and outcome in relation to commodity and services in a period of time, GDP is one of main economic characteristic since it gives demonstration of the economic activity, higher value of GDP indicates better economic activity while lower value indicates the contrary, expressed in billions of national currency units An important statistic that indicates whether an economy is expanding or contracting, GDP can be tracked over long spans of time and used in measuring a nation's economic growth or decline, as well as in determining if an economy is in recession (generally defined as two consecutive quarters of negative GDP growth). [11]Noted that GDP is still one of the most important indicators of macroeconomic statistics, it is an effective tool to make people understand and grasp a national (or regional) macroeconomic status, and it is a scientific and effective method to inspect economic policy and assess the important comprehensive index of economic situation.

1.1.1. Overview of Value added

The various sectors of any economy have a contribution to the development of any economy. This is to say that no matter how small the contribution of any sector to the national income of that economy, it adds up to the aggregate income of the economy and thus contributing directly or indirectly to the gross domestic earnings of such economy [12].The value-added measure of GDP adds together the value of output produced by each of the productive sectors in the economy using the concept of value added. Value added is simply the increase in the value of goods or services as a result of the production process. Our study chose the three among the many economic sectors that contribute towards the growth of GDP and these include Service sector, Agriculture and Manufacturing. This section briefly discusses the sectors.

1.1.2. Service Sector

The tertiary sector or service sector is the third of the three economic sectors of the three-sector theory. The others are the secondary sector (approximately the same as manufacturing), and the primary sector (raw materials). The service sector consists of the parts of the economy, i.e. activities where people offer their knowledge and time to improve productivity, performance, potential, and sustainability, which is termed as affective labor. The basic characteristic of this sector is the production of services instead of end products. Services (also known as "intangible goods") include attention, advice, access, experience, and discussion. The production of information has long been regarded as a service, but some economists now attribute it to a fourth sector, the quaternary sector.

The tertiary sector of industry involves the provision of services to other businesses as well as final consumers. Services may involve the transport, distribution and sale of goods from producer to a consumer, as May happen in wholesaling and retailing, or may involve the provision of a service, such as in pest control or entertainment. The goods may be transformed in the process of providing the service, as happens in the restaurant industry. However, the focus is on people interacting with people and serving the customer rather than transforming physical goods.

1.1.3. Manufacturing Sector

Manufacturing sector refers to those sectors which involve in the manufacturing and processing of items and indulge in either creation of new commodities or in value addition. The manufacturing sector is considered to be one of the prominent sectors for the revitalization of the economy in the strategy for Zambia's socio-economic development and poverty reduction. The country's manufacturing sector comprises of companies in food processing, beverages, textiles, leisure and sporting equipment. The activities majorly include the smelting and refining of copper and other metals and metal products, petroleum refining, the production of fertilizers, chemicals, explosives, cement, tobacco products and textiles. The manufacturing industry accounts for a significant share of the industrial sector in developed countries. The final products can either serve as a finished good for sale to customers or as intermediate goods used in the production process. According to [13] Zambia has a relatively diversified manufacturing sector, which is concentrated in the food, beverages and tobacco sub-sector, accounting for 63 per cent of manufacturing sector activities. Wood and wood products, which is the second dominant sub-sector, makes up 11 per cent of the manufacturing sector. It is followed by the chemicals, rubber and plastics products, which constitutes 9 per cent of the sector.

1.1.4. Agriculture Sector

The Agriculture sector is a sector which comprises establishments primarily engaged in growing crops, raising animals, and harvesting fish and other animals from a farm, ranch, or their natural habitats. Additionally agricultural yield primarily depends on environmental factors such as rainfall, temperature and geographical topology of the particular region. These factors along with some other influence the crop cultivation [4].

2 Data Collection

The study made use of secondary data obtained from World Bank national accounts data, and OECD National Accounts data files which is available on their website.26 years interval (1990 – 2015) [14].Contribution of the identified sectors of the nation's GDP were also obtained from the same source of the same year interval. The three (3) identified sectors include: agriculture, manufacturing, and service value added. The dataset used was loaded in R for analysis, the dataset imported as CSV file, below is the dataset.

Year	**GDP-growth**	**Services-value-added**	**Manufacturing-value-added**	**Agriculture-value-added**
1990	-0.481072127	28.12360801	36.0610909	20.60479675
1991	-0.036133391	31.56585227	36.74848648	17.42768673
1992	-1.730922033	27.21993643	37.16310495	23.81226939
1993	6.797273924	24.00501499	27.94480912	34.10152067
1994	-8.625441835	43.63323144	10.42293014	13.56260604
1995	2.897668709	45.03132238	10.42252367	16.00057803
1996	6.218546514	47.18508198	12.39794018	15.10412944
1997	3.814007573	47.21561167	12.29002026	15.91910159
1998	-0.385746225	50.96601704	12.24400329	18.03400006
1999	4.650189789	53.66768149	11.53670967	20.55077743
2000	3.89732287	55.4134235	10.69943865	18.2786512
2001	5.316868319	55.80267884	10.3217762	17.64738381
2002	4.506014409	55.89433695	10.77533324	17.2675372
2003	6.944973873	55.57070429	11.342165	17.45460527
2004	7.032395131	54.18088113	11.17279892	17.29544049
2005	7.23559899	54.15917557	10.88597265	16.13852965
2006	7.903694512	52.14726684	10.32499946	14.4936135
2007	8.352436224	51.88007276	9.489240695	13.23450525
2008	7.773895812	53.59948565	9.241890106	12.52152896
2009	9.220348444	55.23794946	9.305427104	12.3781784
2010	10.29820581	55.89983928	8.022200348	9.972843012
2011	5.564624717	53.33181109	7.955301724	10.21165686
2012	7.597616969	56.27467003	7.48550623	9.861323998
2013	5.059376378	56.52608605	6.587368149	8.759118438
2014	4.695826373	57.39439898	7.314753999	7.271416345
2015	2.91988111	59.43534303	7.929997344	5.250774058

Fig.8.csv file

3 Data Mining Process

Data mining techniques and methods are used in the main related disciplines and technologies from the following areas: Statistical Methods, Decision Tree, Neural Network, Genetic Algorithm and Fuzzy Set, in [1]. The data mining techniques represent such a tool that solves different types of problems from banking and finance domains, by finding patterns, correlations, rules sets, causalities etc., and helps the human analyst in the process of analysis and prediction of some financial tasks evolution [15], furthermore, some data mining packages offer statistical methods, such as principal components, logistic regression, correspondence analysis etc., for financial predictions.

3.1 Statistical Methods

In data mining this technique often involves a certain degree of statistical process, as data sample and modeling to determine assumptions and error control. Including descriptive statistics, probability theory, regression analysis, time series, including many of the statistical methods, data mining plays an important role. For this study the data mining process was conducted with regression, time series/trend analysis and correlation techniques on the dataset.

3.2 Regression and Value Prediction Technique

Regression according to [16]is a data mining (machine learning) technique used to fit an equation to a dataset. [16] Adds that a regression algorithm estimates the value of the target (response) as a function of the predictors for each case in the build data. These relationships between predictors and target are summarized in a model, which can then be applied to a different data set in which the target values are unknown. In a related study [17] says that regression models make use of relationships between the variable of interest and one or more related predictor variables. Sometimes regression models are called causal forecasting models, because the predictor variables are assumed to describe the forces that cause or drive the observed values of the variable of interest. Regression analysis is a statistical technique for modeling and investigating the relationships between an outcome or response variable and one or more predictor or regressor variables. The end result of a regression analysis study is often to generate a model that can be used to forecast or predict future values of the response variable given specified values of the predictor variables .In this study however we used multiple regression which has been outlined below, multiple regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data [18].

3.2.1 Model Specification

In order to produce an observed study we established a functional relationship for Gross Domestic Product (GDP) as a proxy. As a result of this research, the functional model measurement was represented mathematically as follows:

GDP= f (SVA, MVA, AVA, μ)…………………………….. (a)

Where;

GDP – Gross Domestic Product Growth Rate

SVA – Service Value Added

MVA – Manufacturing Value Added

AVA – Agriculture Value Added

μ- error term

In order to understand the relationship and significance of the variables above, the econometric model was expressed as follows from (a) above:

$GDP = \beta_0 + \beta_1 SVA + \beta_2 MVA + \beta_3 AVA + \mu$……………………… (b)

Where;

GDP – Gross Domestic Product Growth Rate

SVA – Service Value Added

MVA – Manufacturing Value Added

AVA – Agriculture Value Added

μ- error term $\beta 0$= Constant term and

$\beta 1$= the coefficient of the independent variable (SVA) which $\beta_1 > 0$

$\beta 2$= the coefficient of the independent variable (MVA) which $\beta_2 > 0$

$\beta 3$= the coefficient of the independent variable (AVA) which $\beta_3 > 0$

3.2.2 Identification of Variables

- The independent variables are Service Value Added, Manufacturing Value Added, and Agriculture Value Added.
- The dependent variable is Gross Domestic Product Growth Rate

Summary of the output is shown below

Table 1 Summary of Output

Regression					
Multiple R	0.53491				
R Square	0.286129				
Adjusted R Square	0.188782				
Standard Error	3.688029				
Observations	26				
	df	*SS*	*MS*	*F*	*Significance F*
Regression	3	119.9368	39.97894	2.939291	0.055642
Residual	22	299.2343	13.60156		

Table 2 Regression Summary

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	-14.4138	13.39847	-1.07578	0.293682
Services value added	0.303039	0.198561	1.526175	0.141216
Manufacturing value added	0.012526	0.199203	0.06288	0.95043
Agriculture value added	0.246642	0.182411	1.352122	0.190074

3.2.3 Interpretation of Results and Predictions

Constant Interpretation

The intercept value of -14.4138 was found, mechanically interpreted, means that if the values of SVA,MVA, and AVA rate were fixed at zero, the GDP growth rate would reduce by 14.4138 per year. This implies that, GDP growth rate will still decrease by 14.4138 when service value added, manufacturing value added and agriculture value added involved are held constant.

Coefficients Interpretation:

1. **Service Value Added**

0.303039 is the partial regression coefficient of SVA and tells us that with the influence of MVA and AVA held constant, as SVA increases by 1 percent, GDP growth rate in a year will increase by 0.303039. The coefficient was positive and statistically significantly with t-test value of 1.526175 at 0.05 level of significance. This means that, if service value added increases by 1 percent, GDP growth rate will increase by 30.3 percent. The service sector is a very promising sector to improve Zambian GDP growth rate, thus the government need to improve and invest more in the service industry. This implies that, if the service sector increases it value added by at least 1 percent, then the Zambian GDP growth rate is likely to increase by more than 30 percent

2. **Manufacturing Value Added**

In the model, the partial regression coefficient of MVA was found to be 0.012526 which implies that for every 1 percent increase in manufactured goods, GDP growth rate will increases by 0.012526. The coefficient was positive and statistically significantly with t-test value of 0.199203 at 0.05 level of significance. This economically means that, if manufacturing industry continues to produce goods for domestic and foreign consumption then GDP growth rate will continue increasing by 1.2 percent. This means that, government has a lot to do in manufacturing industry because it is not contributing more to GDP according to its potential. If manufacturing industry is heavily funded, then it will bring more of foreign exchange to the country which will make the country to have surplus trade balance and in return boosting the national income (GDP). As a result, if manufacturing industry increases it value added by at least 1 percent, then the Zambian GDP growth rate is likely to increase by more than 1.2 percent

3. **Agriculture Value Added**

The partial coefficient of AVA was found to be 0.246642 and tells us that holding the influence of SVA, and MVA coefficient, constant, GDP growth rate in a year will increase by 24.6642 percent at 0.05 level of significance. Economically, it implies that GDP growth rate will still increase by 24.6642 percentage rate. The government of Zambia should invest more or improve budget allocation to agriculture sector as it is a viable sector which can improve the GDP growth rate. If agriculture industry is well financed through the processing of value added on raw materials, then more of foreign exchange will be brought to the country and increase the GDP growth rate. This shows that, if agriculture sector increases its value additional by more than 1 percent, the Zambian GDP growth rate will likely to increase by 24.6 percent or more than 24.6 percent.

4. **Coefficient of Variation R-Squared (*R^2*)**

The coefficient of variation for this econometric model was found to be 0.286129. This implies that 27 percent of the disparity on total trips traveled was attributed to the variation on service value added, manufacturing value added and agriculture value added and the remaining 73 percent of the variables in the model are not explained. The R^2 is greater than 0.2 as a result the fitted regression line is of good fit.

5. **F-test and its significance:**

The F-test shows that the relation of the disparity in the regression to the ratio of the variation of the residual or errors was found as 2.939291. This value found was greater than the tabulated F-value of 0.055642 at 0.05 level of significance. This implies that the econometric model was significantly correct.

4 Trend Analysis using time series

Time series data type, also called chronological series or simply time series represent results of measurements made on the characteristics of a unit of population studied, over time, at successive moments of its evolution in some time intervals [19]. Using this technique we wanted to find out the trend of gross domestic product with respect to service, agriculture and manufacturing sector for the past decade. Below is plotted graph showing the trend with respect to the aforementioned variables.

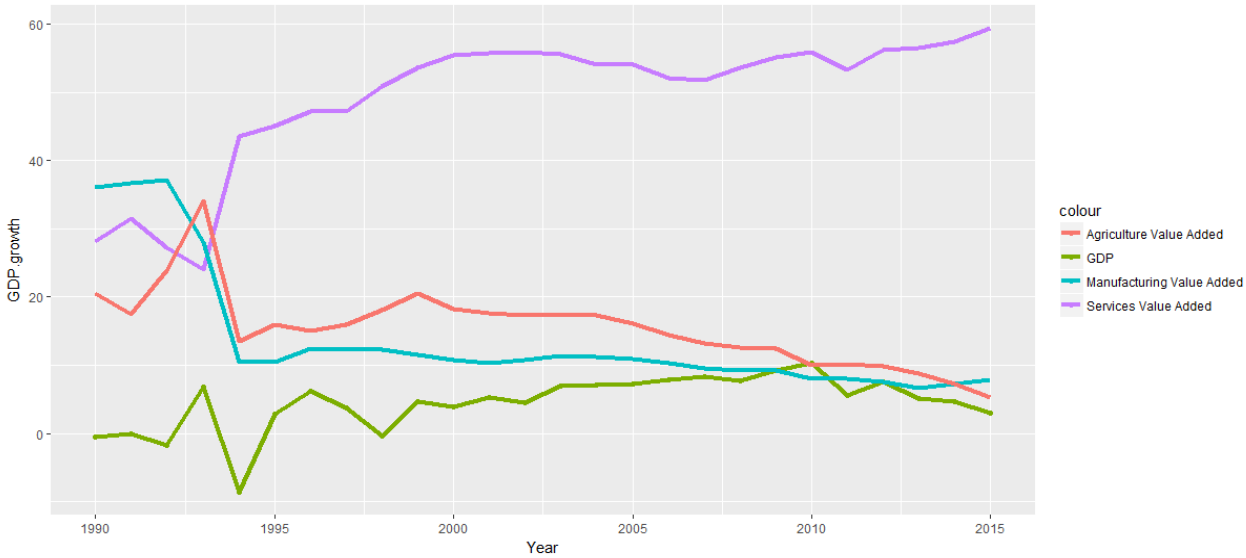

Fig. 9. GDP Growth, Service Value added, Manufacturing Value Added and Agriculture Value Added trend analysis

4.1. Interpretation and Results Prediction

4.1.1. Service Value Added

The service sector consists of the parts of the economy, i.e. activities where people offer their knowledge and time to improve productivity, performance, potential, and sustainability, which is termed as affective labor or a system supplying a public need such as transport, communications, or utilities such as electricity and water etc. The Zambian service sector has the potential to boost the Zambian GDP growth rate, because it has kept increasing for the past two decade. The Zambian government needs to start providing and training skilled human capital and improve on a number of services within the economy as it has shown potential to make GDP grow. The government can do more in the service sector to have a desirable GDP growth rate.

4.1.2. Manufacturing Value Added

The manufacturing sector in Zambia accounts for about 11 percent of the country's Gross Domestic Product (GDP) and has been growing at an average annual growth rate of three (3) percent in the last three years. Growth in the sector is largely driven by the agro processing (food and beverages), textiles and leather subsectors. Secondary processing of metals in another main activity in the sector, including the smelting and refining of copper, and this has led to the manufacturing of metal products. Fertilizers, chemicals, explosives and construction materials such as cement are also produced in the sector. Other activities include wood products and paper products. The manufacturing activities in the country are undertaken by the private sector with government playing a proactive role. The sector is of vital importance in relation to the country' macroeconomic strategy for encouraging broad based economic growth. In this regard, the Government has put in place measures to support manufacturing activities, such as the establishment of Multi-Facility Economic Zones (MFEZs) and Industrial Parks (these are industrial areas for both export orientated and domestic orientated industries, with the necessary support infrastructure installed), and provision of sector-specific investment incentives. Government also promotes small and medium enterprises in rural and urban areas so as to enhance labor intensive light manufacturing activities in these areas. In order for Zambia manufacturing industry to improve or growth, value additional should the key factor to implement as most of the goods are been sold in raw form or in unfinished products. And if the Zambian government can have a deliberate policy to have the manufacturing industry add value to unfinished products before exporting them as to enjoy the full value price for the product which in turn will improve the GDP growth rate.

4.1.3. Agriculture Value Added

For the past two decades, the Zambian agriculture sector has not been doing well, as the result, the productivity has continued to decline because of a number of factors such as: over dependence of small scale farmers on farmers input support programme (FISP), government policies, climate and over dependence on copper. The Zambian government has continued to invest in the agriculture sector, according to Ministry of Agriculture and Livestock acting permanent secretary, who also praised the industry's role in the economy in a recent statement issued in Lusaka. The performance of the industry in the just ended year was significant with the growth of GDP about seven per cent. The Zambian government has announced it will implement plans to boost crop and livestock production to help further strengthen the agriculture sector. It said it would invest in extension services, irrigation, and research and development in order to improve crop yield. The government also claimed it was working to increase local production of key inputs such as fertilizers, seeds and pesticides, along with restocking the livestock sector with a view to improve its performance. If the government can continue boosting the agriculture sector, it can increase the gross domestic product by a bigger percentage, because this sector has a potential to boost the Zambian economy and a lot has to be done to improve the value creation or adding of value to the raw material as to boost the value creation in agriculture sector which in turn will boost the Zambian GDP.

4.1.4. GDP Growth

Gross domestic product (GDP) is the monetary value of all the finished goods and services produced within a country's borders in a specific period of time. Though GDP is usually calculated on an annual basis, it can be calculated on a quarterly basis as well. GDP includes all private and public consumption, government outlays, investments, private inventories, paid-in construction costs and the foreign balance of trade (exports are added, imports are subtracted). From the graph, we can deduce that, they have been some reduction in Zambian GDP growth rate for the past two decades. The Zambian GDP growth rate will continue to decrease, unless the Zambia government improves on the service, agriculture and manufacturing sector. In order to improve the Zambian GDP, the government should focus on service, agriculture and manufacturing sector.

5 Correlation Technique

Correlation technique was chosen to help us find out the relationship among the variables (service, agriculture and manufacturing sector) with respect to GDP as a controlling variable.

The table below shows correlation relationships amongst the variables

Table 3. Variable Correlation Relationship

	Year	GDP growth	Services Value added	Manufacturing value added	Agriculture value added
Year	1	0.4849692	0.7734849	-0.5000892	-0.2820188

GDP growth	0.484969	1	0.4491033	-0.4298382	-0.1623941
Services Value added	0.773485	0.4491033	1	-0.7928582	-0.5073032
Manufacturing value added	-0.50009	-0.4298382	-0.7928582	1	0.5588934
Agriculture value added	-0.28202	-0.1623941	-0.5073032	0.5588934	1

Table 4. Correlation Relationship among service, manufacturing and agriculture

Correlations Summary showing the relation among service, manufacturing and Agriculture variables					
Control Variables			Service	Manufacturing	Agriculture
GDP	Service	Correlation	1.000000	-0.792**	-0.507**
	Manufacturing	Correlation	-0.792**	1.000000	0.558**
	Agriculture	Correlation	-0.507**	0.558**	1.000000
**. Correlation is significant at 0.01 level					

In statistics, the correlation coefficient *r* measures the strength and direction of a linear relationship between two variables . The value of *r* is always between +1 and –1. To interpret its value, we follow values the correlation *r* is closest to:

- Exactly –1. A perfect downhill (negative) linear relationship
- –0.70. A strong downhill (negative) linear relationship
- –0.50. A moderate downhill (negative) relationship
- –0.30. A weak downhill (negative) linear relationship
- 0. No linear relationship
- +0.30. A weak uphill (positive) linear relationship
- +0.50. A moderate uphill (positive) relationship
- +0.70. A strong uphill (positive) linear relationship
- Exactly +1. A perfect uphill (positive) linear relationship

5.1. Interpretation of Results and Prediction

5.1.1 Service Value Added

There is a strong negative correlation (-0.792) between service value added and manufacturing, and at the same time, there is a weak negative correlation (-0.507**) between service value added and agriculture value added.

5.1.2 Manufacturing Value Added

There is a strong negative correlation (-0.792) between service value added and manufacturing, and at the same time, there is a moderate correlation (0.558**) between manufacturing value added and agriculture value added.

5.1.3 Agriculture Value Added

There is a negative correlation (-0.507**) between agriculture value added and service value added, and at the same time, there is a moderate correlation (0.558) between manufacturing value added and agriculture value added.

From the above correlation analysis, manufacturing and agriculture value added seems to correlate and it's very possible that if these two sectors are boosted through value addition, the Zambian GDP will grow at a steady increasing rate. These two sectors are viable and government has to pump in a lot of funds (investment) as they will positively increase Zambian GDP growth rate.

6. Limitation of the techniques

6.1. Multiple Regression Model

Multiple Regression Model has a problem of problem of multicollinearity (One or more variables explaining the same factor i.e. Mother and a lady) as it may exist between or among variables. In such an event, one or more variables should be eliminated to reduce or eliminate multicollinearity. As a result, it helps to check or eliminate the highly correlated independent variables from the analysis, recognizing that the two variables essentially are measuring the same factors and there is no need of having both variables. Shanta Khumari in [12] also noted that multicollinearity causes major interpretative problems in regression analysis, such as wrong sign problem, produces unstable and inconsistent estimates of parameter, insignificant regression coefficients where in fact it is significant and it is thus very essential to investigate and detect the presence to reduce the destructive effects of multicollinearity. Furthermore, it is very difficult to explain the coefficient of variables in a multiple regression model as it may not be simplified in layman's language.

6.2. Time series analysis

Time series analysis helps to show the trends about the variables over a period of time (It shows the up and downs about a variable). The major limitation about time series analysis is that, it does not show the statistical figures about the variables and which makes it difficult for the researchers to forecast the future behaviour for the variables.

6.3. Correlation Analysis

Correlation analysis is an essential analysis when one wants to see the relationship between or among variables that are being analysed (It ranges between -1 to +1) and it helps to see how variables correlate among each other. The challenge which this techniques brings during correlation analysis is that, the analysis only shows the correlation among variables and makes it difficult for the researchers to predict the future behavior of variables based on correlation of variables. The analysis does not show the coefficients of variables or the statistical figure which may add more value or significance on the analysis.

Conclusions and Recommendations

7 Conclusion

In this paper certain data mining techniques were adopted to analyze the data that shows relevance with desired attributes. Regression technique was adopted to help us find out the influence of Agriculture, Service and Manufacturing value added on the performance of gross domestic product (GDP). Trend and time series technique was applied to the data to help us find out what trend of GDP with respect to service, agriculture and manufacturing sector for the past decade has been. Finally Correlation was also used to help us analyze the relationship among the variables (service, agriculture and manufacturing sector) with respect to GDP as a controlling variable. Many data mining techniques were however not employed on the GDP dataset because the study had set out objectives that were only going to be solved with the aforementioned techniques. From the three techniques analyzed, service value added variable was the most prominent variable which showed the strong influence on GDP growth rate (Under multiple regression service value added had highest positive coefficient, under time series – it had a strong positive correlation and under correlation analysis – GDP growth rate and service value added were more correlated than any other variable). From the three techniques used, regression analysis was easy to use for predicament influence of independent variables on dependent variable as compared to the other two techniques used. There are however other variables which affect the GDP growth rate apart from the three independent variables which were analyzed. The researchers had to pick three independent variables as to reduce the level of complication in analysis and making it simpler for the reader to understand. Future work however will consider utilizing other data mining techniques that will help us gain a deeper understanding and mining of GDP data on a different perspective. Overall as scholars we have come to appreciate data-mining because it helps in making informed decisions, derive new knowledge, analyze the data and make predictions that in turn if well utilized as far as the growth of the economy is concerned can greatly contribute positively and certain measures can be put in place to help the economy grow for the betterment of our nation Zambia.

8Recommendations

Agriculture Sector

The limited financial and non-financial support available has curtailed the overall growth of some nonfarming sub-sectors of agriculture, which has reduced their overall contribution to wealth (income) and employment creation. Within these sectors as well, it is important to invest in forward and backward integration to increase the value of products exported and the number of jobs created. Furthermore, for certain sub-sectors, such as livestock and wood processing, there is a need to develop the skills endowment levels of the labor force [13]. Inorder to boost value addition in the agriculture sector, following should be considered by government:

i. Review the national agriculture policy to prioritize livestock development and fish farming by supporting both production and market infrastructure for value added production within these sub-sectors.
ii. Reform rural agricultural co-operatives into business-oriented entities focused on adding value to agricultural commodities. In particular, this should be done by pooling resources to upgrade or purchase new machinery and technology.
iii. Increase investments to expand rural irrigation infrastructure to lessen smallholder farmer dependency on rain-fed agriculture.
iv. Strengthen agricultural extension and veterinary services. Also, link livestock development research to the smallholder sector to help control animal disease outbreaks and enhance the productivity of this sector.
v. Resolve institutional issues, such as improving the collection levels and supply of quality raw hides and skins, and ensuring that producers earn competitive prices to stimulate growth and counter the rampant smuggling of these vital agricultural inputs.
vi. Improve inter-ministerial coordination across agriculture,

industry, trade, and employment, to ensure the development of policies and a regulatory framework that is conducive for sectoral growth and development.

vii. Accelerate the implementation of land reforms, especially under the Customary Tenure. These reforms are critical to guaranteeing land rights which can, in turn, allow for capital investment in land and provision of credit.

viii. Set aside funds or provide an enabling regulatory environment to attract investments in suitable R&D that would enable Zambian manufacturers to gain increasing access to competitive global markets.

ix. Prioritize entrepreneurship development, technical skills, and management training across this sector to promote productivity gains and accelerate the commercialization of smallholder agriculture.

x. Realign the agricultural sector budget to give equal emphasis to crop agriculture, horticulture, and livestock development.

xi. Promote sustainable exploitation of fisheries resources and increased fish production [20].

Manufacturing Sector

Under the manufacturing sector for value addition, there is need for strengthening the sector and putting in place certain measures and viable strategies such as the ones outlined below:

i. Accelerating the policy and enactment of the commerce, trade and industrial related policies to prioritize wealth and employment creation through value added manufacturing.

ii. Parallel diversification of the mining activities towards industrial minerals such as iron ore and steel milling, to introduce a local capital goods sector. The policy goal should be to hasten the growth of a vibrant and competitive iron and steel industry to become the anchor of industrial policy. This should support the growth of automotive components, and medium and heavy industry commercial vehicles to support the growth and expansion of mining and mineral beneficiation.

iii. Embark on strategic partnerships and joint ventures with public and private, as well as foreign and domestic investors to provide skills, technical expertise, and technology to target strategic sectors and integrate them into the global value chains.

iv. Prioritize research and development and provide smarter subsidies to accelerate the development of the leather and leather products subsector.

v. Establish specialized research and training Institutes and technical colleges to provide the necessary technology and local capability to support the development of a competitive agro-industry, textile and clothing, iron and steel, and leather and leather products industry.

vi. Prioritize the development of the forest and forestry sector through consistent funding of tree planting to provide the much-needed raw material to the wood and furniture, and paper and paper products, sub-sectors.

vii. Invest in the expansion of green and energy-saving industries, such as hydroelectricity, solar water heating, concentrated solar power, and improvements in energy efficiency [13].

viii. Promote entrepreneurship development and training at all levels of the education system.

Service sector

The Zambian government needs to start providing and training skilled human capital and improve on a number of services within the economy as it has shown potential to make GDP grow. The government can do more in the service sector to have a desirable GDP growth rate. Additionally Government needs to develop and adopt a comprehensive strategy that will see the creation of an environment for new industries to spring up, add-value and diversify the products for export

References

[1] D. Bhardwaj, "ANALYSIS OF DATA MINING TRENDS, APPLICATIONS, BENEFITS AND ISSUES," International Journal of Computer Science and Communication Engineering, vol. 5, no. 1, pp. 1-5, 2016.
[2] M. R. Bharati, "DATA MINING TECHNIQUES AND APPLICATIONS," Indian Journal of Computer Science and Engineering, vol. 1, no. 4, pp. 301-305, 2011.
[3] Z. M. Fathimath and A. Geetha, "Analysis of Data Mining Techniques and its Applications," International Journal of Computer Applications, vol. 140, no. 3, pp. 6-14, 2016.
[4] A. B. Vinayak, P. A. Prachi, S. P. Ravina, S. P. Sonal, U. N. Tejaswini and M. J. Anaya, "Analysis and Prediction in the Agriculture data using data mining techniques," International Journal of Research In Science & Engineering, pp. 386-393, 2017.
[5] L. Yang and H. Bin, "The Establishment and Analysis of Time Series Model of Per Capita GDP in Yunnan, China," pp. 1-3, 2010.
[6] A. Naveed, A. Kartheek, P. Tapan and V. Meghana, "Predicting Economic Recession using Data Mining Techniques".
[7] R. K. Alex, "The Factors Affecting Gross Domestic Product (GDP) in Developing Countries: The Case of Tanzania," European Journal of Business and Management, vol. 5, no. 4, pp. 148-158, 2013.
[8] M. Gregory, Principles of Economics, Cengage Learning, 2014, p. 880.
[9] R. Barbara and Z. Andrea, "EXCESS MONEY GROWTH AND INFLATION DYNAMICS," pp. 1-42, may 2007.
[10] S. Siti, H. ,. The and A. , "Prediction System of Economic Crisis in Indonesia Using Time Series Analysis and System Dynamic Optimized by Genetic Algorithm," International Conference on System Engineering and Technology, pp. 1-6, 2012.
[11] G. LEI and Z. HUI-BEN, "THE ANALYSIS OF AFFECTING GDP GROWTH FACTORS BASED ON EVIEWS ECONOMETRIC MODEL," 2013.
[12] V. E. Ikenna, Olugbenga and O. K, "Sectoral Contribution to Nigeria's Gross Domestic Product (GDP) Growth Rate: A Study of Multicollinearity in Aggregated Time Series Data," Journal of Scientific Research & Reports, vol. 11, no. 1, pp. 1-13, 2016.
[13] ZHDR, "Zambia Human Development Report 2016," UNDP, Lusaka, 2016.
[14] a. O. N. A. d. f. World Bank national accounts data, "GDP growth (annual %)," [Online]. Available: http://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?locations=ZM. [Accessed May 2017].
[15] M.Cărbureanu, "The Analysis of Currency Exchange Rate Evolution using a Data Mining Technique," Economic Sciences, vol. LXIII, no. 3, pp. 105-112, 2011.
[16] G. Swati, "A Regression Modeling Technique on Data Mining," International Journal of Computer Applications, vol. 116, no. 9, pp. 27-29, 2015.
[17] S. SAIGAL and D. MEHROTRA, "PERFORMANCE COMPARISON OF TIME SERIES DATA USING PREDICTIVE DATA MINING TECHNIQUES," Advances in Information Mining, vol. 4, no. 1, pp. 57-66, 2012.

[18] C. Hao and Y. Lixin, "Using Multi-regression to Analyze and Predict Road Traffic Safety Level in China," in The 3rd International Conference on Transportation Information and Safety, Wuhan,, 2013.
[19] J. Popeanga, "Data Mining Smart Energy Time Series," Database Systems Journal, vol. 1, no. 1, pp. 14-22, 2015.
[20] M. o. F. a. N. Planning, "SIXTH NATIONAL DEVELOPMENT PLAN 2011-2015," Ministry of Finance and National Planning, Lusaka, 2011.
[21] G. Matjaž and J. Krivec, "Demographic Analysis of Fertility Using Data Mining Tools," Informatica , pp. 147-156, 2008.

Assessment of the Effects of Electricity consumption on the Economy using Granger Causality: Zambian Case

Douglas KUNDA, Mumbi CHISIMBA
Mulungushi University, School of Science Engineering and Technology, P.O Box 80415, Kabwe, Zambia
Email: dkunda@mu.edu.zm* and chishimba.mumbi@gmail.com
Phone Number: +260965883588*

Electricity consumption in developing countries such as Zambia continues to grow as the economy grows. As a result, it is important to study how the rate of electricity consumption affects the economy of a country. For this study, the economic variables that were used are the Gross Domestic Product and the Consumer Price index. The results from this study are that there is a unidirectional relationship between electricity consumption and the consumer price index where the rate of electricity consumption Granger causes the consumer price index. The study also showed that there is no causal relationship between electricity and GDP and that there was no causal relationship between electricity consumption and the Consumer Price Index.
Keywords: *Electricity consumption; GDP; CPI; Granger Causality; Zambia*

1 Introduction

Interest and study in the topic of the relationship between electricity consumption and other fields such as economic growth and demographics is a well-studied field in economic literature [1]. This area of study has seen increased interest since the oil crisis in the 1970s and the energy crisis due to peak oil in the 2000s which put pressure on countries to conserve energy and find ways of efficiently using the electricity that is available [2]. This, however is not always the case in developing countries such as Zambia which need the extra energy usage in order to develop their economies [2]. As such, it is important to properly study the relationship between the rate of electricity consumption and the economy. Various studies on the casual relationship between electricity consumption and the economy have produced various results as shown in Table **5**. The country of Zambia has seen an increase in electricity production as shown in Fig.10.

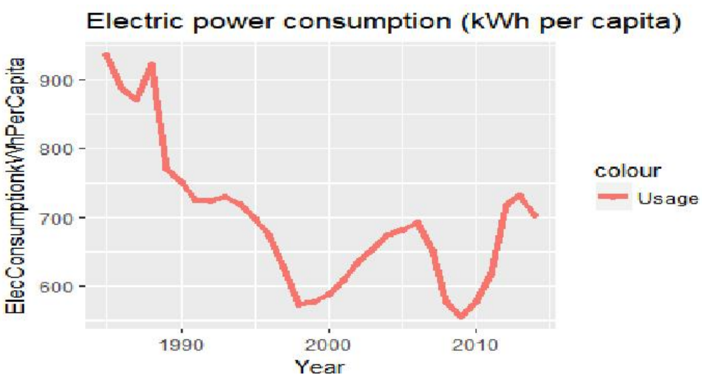

Fig.10. Electricity production in Zambia

Zambia has also seen an increase in electricity demand has led the national electricity supply company, ZESCO (Zambia Electricity Supply Corporation), to ration electricity through the practice of load shedding [3].

In order to best understand the relationship between electricity consumption or production and the economy, there is a need to understand the causal relationship between the two factors. There are three possible causal relationship categorizations namely (1) no causality, (2) unidirectional causality and (3) bi-directional causality between energy consumption and economic growth [1]. Unidirectional relationships can be further divided into two aspects namely (a) energy consumption causes economic growth and (b) economic growth causes energy consumption [1].

2. Theoretical background

The energy sector in Zambia is diverse in that the Zambian energy sector includes sources such as electricity, petroleum, coal, biomass, and renewable energy sources and from these sources, petroleum is the only one which is wholly imported in the country [4]. Also, as of 2014, there was a growth of demand of electricity and other forms of electricity for a rate of about 3% per annum mostly due to the increase in economic activity in the country. Overall, as of 2014, hydro-electricity is the most important energy source in the country after wood fuel which contributes about 10% of the national energy supply [4]. Though Zambia possesses about 40% of the water resources in the Southern African Development Community, Zambia has about 6,000 MW unexplored hydro power potential while only 2, 177 MW of power has been developed [4]. Fig.11 shows the installed generation capacity of electricity in Zambia.

No	Power Station	Installed Capacity	Type of Generation	Operator
1	Kafue Gorge	990	Hydro	
2	Kariba North Bank	1,080	Hydro	ZESCO
3	Victoria Falls	108	Hydro	
4	Lusemfwa and Mulungushi	56	Hydro	Lusemfwa Hydro Corp.
5	Small Hydros - combined	25	Hydro	ZESCO
6	Isolated Generation	8	Diesel	
7	Gas Turbine (stand by)	80	Diesel	Copperbelt Energy Corp.
	Total Installed Capacity	**2,177**		

Fig.11. Installed Generation Capacity in Zambia (Mega Watts) [4]

Aside from hydro-electricity, Zambia also makes relatively high use of petroleum products in Zambia. In total, petroleum contributed, as of 2014, an estimated 9% of the national energy requirements is imported and plays a crucial role in agriculture, transport and mining [4]. The petroleum is imported in the form of crude oil through the Tanzania-Zambia Mafuta oil pipeline and is refined at the Indeni oil refinery in Ndola city. Just like electricity, the demand for petroleum in Zambia has also increased in recent years. As of 2014, the demand for petroleum was at around 52 million litres per month. <FIGURE> shows the demand for petroleum products in Zambia as of 2014.

	Type of Petroleum Product	Average monthly consumption (liters)
1	Petrol Premium	12,000,000
2	Diesel / Gas Oil	30,000,000
3	Liquefied Petroleum Gas	190,000
4	JET-A-1	2,900,000
5	Heavy Fuel Oil	5,800,000
6	Kerosene	918,000
	TOTAL	**51,808,000**

Fig.12. Demand for Petroleum Products in Zambia [4]

Another source of energy in Zambia is that of bio-fuels. Though Zambia does not possess the capacity to produce bio-fuels, there are a number of areas where bio-fuels are used. Efforts such as production of bio-ethanol from molasses are being pursued though the though it is not being blended with petrol [4]. The Ministry of Energy and Water Development in Zambia estimated that an estimated 84 million litres of bio diesel and approximately 40 million litres of bio-ethanol are required by the country per annum as shown in Fig.13 and Fig.14.

Year	Diesel Sales Volume	5% Blend	10% Blend	15% Blend	20% Blend
Millions of Litres					
2008	379.10	18.96	37.91	56.87	75.82
2009	398.10	19.90	39.81	59.71	79.61
2010	417.96	20.90	41.80	62.69	83.59
2011	436.80	21.96	43.70	64.18	87.42
2012	452.18	22.90	45.80	65.90	90.21

Fig.13. Estimated Biodiesel Demand in Zambia (2008 to 2012) [4]

Year	Petro Sales Volume	10% Blend	15% Blend	20% Blend
Million Litres				
2006	176.15	17.62	26.42	35.23
2007	184.96	18.50	27.74	36.99
2008	194.21	19.42	29.13	38.84
2009	203.92	20.40	30.59	40.78
2010	203.92	21.41	32.12	42.82

Fig.14. Estimated Bio-ethanol Demand in Zambia (2006 to 2010) [4]

Table 5. Studies on the Relationship between Electricity and the economy

Study	Countries Studied	Variables Used	Method	Result
[2]	Pakistan	Electricity consumption and GDP	The augmented Dickey–Fuller (ADF) and Phillips–Perron (PP) tests	Using annual data for the period 1960–2008, the study finds the presence of unidirectional causality from real economic activity to electricity consumption
[5]	China, Hong Kong, Indonesia, India, Korea, Malaysia, the Philippines, Singapore, Taiwan and Thailand.	GDP and electricity consumption		Electricity conservation policies through both rationalizing the electricity supply efficiency improvement to avoid the wastage of electricity and managing demand side to reduce the electricity consumption without affecting the end-user benefits could be initiated without adverse effect on economic growth. The findings on the long-run relationship indicate that a sufficiently large supply of electricity can ensure that a higher level of economic growth.
[6]	Malaysia	Electricity generation, exports, prices and GDP	Cointegration, Granger causality	Electricity conservation policies, including efficiency improvement measures and demand management policies, which are designed to reduce the wastage of electricity and curtail generation can be implemented without having an adverse effect on Malaysia's economic growth.

[7]	Malawi	GDP and electricity consumption	Granger-causality (GC) and error correction (ECM)	bi-directional causality between kWh and GDP suggesting that kWh and GDP are jointly determined, but one-way causality running from NGDP to kWh.
[8]	China	Real GDP and electricity consumption	Granger causality	Real GDP and electricity consumption for China are cointegrated and there is unidirectional Granger causality running from electricity consumption to real GDP but not vice versa.
[1]	Bangladesh	GDP and electricity consumption	Cointegration and vector error correction model.	There is unidirectional causality from per capita GDP to per capita electricity consumption. However, the per capita electricity consumption does not cause per capita GDP in case of Bangladesh.
[9]	Australia, Austria, Belgium, Canada, Czech Rep, Denmark, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Japan, Korea, Luxembourg, Mexico, Netherlands, New Zealand, Norway, Poland, Portugal, Slovak Rep., Spain, Sweden, Switzerland, Turkey, UK and USA	Electricity consumption–real GDP	Bootstrapped causality testing approach	Evidence in favour of electricity consumption causing real GDP in Australia, Iceland, Italy, the Slovak Republic, the Czech Republic, Korea, Portugal, and the UK. The implication is that electricity conservation policies will negatively impact real GDP in these countries. However, for the rest of the 22 countries our findings suggest that electricity conversation policies will not affect real GDP.
[10]	Russia	Energy Consumption, electricity, and GDP	Granger Causality Test	Both the economic growth and electricity consumption empirically support each other and have a mutual and complementary relationship

A number of studies were performed in order to find out the causal relationship between electricity consumption/production and economic performance. For instance, in a study on Causality relationship between electricity consumption and GDP in Bangladesh [1], P. Mozumder and A. Marathe discovered that there was unidirectional causality from per capita GDP to per capita electricity consumption. However, P. Mozumder and A. Marathe noted that the per capita electricity consumption did not cause per capita GDP in case of Bangladesh. The result was similar to that discovered by F. Jamil and E. Ahmad, Using annual data for the period 1960–2008, which found the presence of unidirectional causality from real economic activity to electricity consumption [2]. Another study that was performed which could be better be related to Zambia is a study performed by C. B. Jumbe, in a study on the Cointegration and causality between electricity consumption and GDPin Malawi. The study showed that there was a bi-directional causality between kWh and GDP suggesting that kWh and GDP are jointly determined, but one-way causality running from NGDP to kWh [7]. In a study on Electricity consumption–real GDP causality nexus in 30 OECD countries, P. K. Narayan and A. Prasad dicovered that there was evidence in favour of electricity consumption causing a negatively impact on real GDP in Australia, Iceland, Italy, the Slovak Republic, the Czech Republic, Korea, Portugal, and the UK. However, findings from the rest of the 22 countries suggest that electricity conversation policies will not affect real GDP [9].

On the electricity conservation front, S.-T. Chen, H.-I. Kuo and C.-C. Chen discovered that electricity conservation policies through both rationalizing the electricity supply efficiency improvement to avoid the wastage of electricity did not significantly affect economic growth [5].

3. Research Model and Hypothesis

3.1. Rate of Electricity Consumption

In electrical engineering, power consumption often refers to the electrical energy over time supplied to operate an electrical appliance.

The first set of hypotheses involve the relationship between Electricity consumption and the Gross Domestic Product at current prices in Billions of US dollars. The null hypothesis (H_0) in this case is that the rate of electricity consumption does not Granger cause the Gross Domestic Product (GDP) of the country. The first alternative hypothesis (H_1) is that the rate of electricity consumption does Granger cause the Gross Domestic Product (GDP) of the country.

H1 Rate of electricity consumption does Granger cause the Gross Domestic Product (GDP) of the country

H2 Rate of electricity consumption does Granger cause the Consumer Price Index

3.2 Gross Domestic Product

According to [11] GDP is one of the measures of national income and output for a given country's economy at a given period of time and adds that the definition of GDP is based on the total market value of all final goods and services produced within the country in a given period of time (normally one year).

The second set of hypotheses involve the relationship between Electricity consumption and the Gross Domestic Product at current prices in Billions of US dollars. The null hypothesis (H_0) in this case is that the Gross Domestic Product (GDP) does not Granger cause the rate of electricity consumption of the country. The first alternative hypothesis (H_1) is that the Gross Domestic Product (GDP) does Granger cause the rate of electricity consumption of the country.

H3 Gross Domestic Product (GDP) does Granger cause the rate of electricity consumption of the country

3.3. Consumer Price Index

The third set of hypotheses involve the relationship between electricity consumption and the Consumer Price Index (CPI). The null hypothesis (H_0) for this set of hypotheses is that the rate of electricity consumption does not Granger cause the Consumer Price Index. The first alternative hypothesis (H_1) is that the rate of electricity consumption does Granger cause the Consumer Price Index.

The final set of hypotheses involve the relationship between electricity consumption and the Consumer Price Index (CPI). The null hypothesis (H_0) for this set of hypotheses is that the Consumer Price Index does not Granger cause the rate of electricity consumption. The first alternative hypothesis (H_1) is that the Consumer Price Index does Granger cause the rate of electricity consumption

H4 Consumer Price Index does Granger cause the rate of electricity consumption

3.4 Methodology

For this study, Electric power consumption (kWh per capita) [12], the consumer price index [13] and the GDP (current US$) [14] in Zambia are used. The data analysed was obtained from the World Bank national accounts data, and OECD National Accounts data files and ranged from the year 1985 to the year 2014.

Fig.15, Fig 16 and Fig.17 shows graphs for the Electric power consumption (kWh per capita), the consumer price index per annum and the GDP (current US$) respectively.

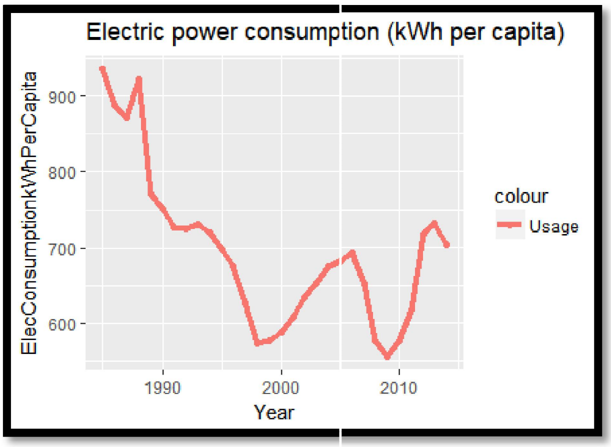

Fig.15. Electricity power consumption (kWh per capita)

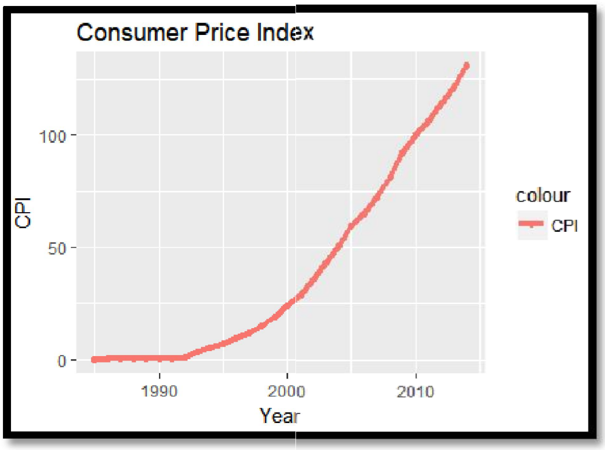

Fig 16. Consumer Price Index

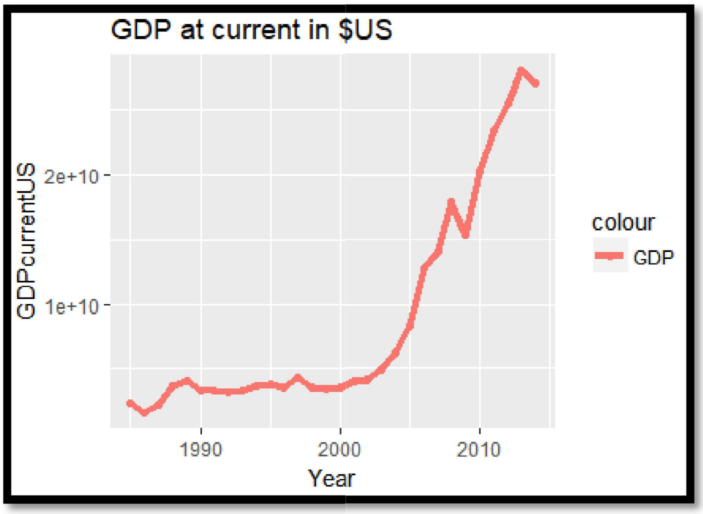

Fig.17.GDP at current in $US

For the data analysis, the Granger causality test to determine whether there is a causal relationship between (1) Electric power consumption (kWh per capita) [12] and the consumer price index, and (2) Electric power consumption (kWh per capita) [12] and the Gross Domestic Product (current US$) [14]. The first step taken in this study was to test the stationarity of all the variables using the Augmented Dickey-Fuller (ADF) test. Granger causality is under normal circumstances tested in the context of linear regression models. The formula for Granger causality is illustrated below:

$$\Delta X_{it} = \beta_{1i} + \sum_{j=1}^{k} \beta_{11ij} \Delta X_{i,t-j} + \sum_{j=1}^{k} \beta_{12ij} \Delta Y_{i,t-j} + \lambda_{1i} \varepsilon_{it-1} + \mu_{2it}$$

$$\Delta Y_{it} = \beta_{1i} + \sum_{j=1}^{k} \beta_{11ij} \Delta Y_{i,t-j} + \sum_{j=1}^{k} \beta_{12ij} \Delta X_{i,t-j} + \lambda_{2i} \varepsilon_{it-1} + \mu_{2it}$$

Fig.18. Granger causality formula **Error! Reference source not found.**

In the formulae above, for the data analysis, two data variables are compared using these formulae e.g. if Electricity consumption and GDP are the variables being compared, X would represent the Electricity consumption and Y would represent the GDP. In this case, the first equation would measure changes in Electricity Consumption and Y would represent changes in the Gross Domestic Product. The Granger causality test is used in this paper the test methodology can be used for heterogeneous panel data models with fixed coefficient**Error! Reference source not found.**. The Granger causality test is also used to test the directionality of the cause I.e. we can find out whether Electricity Consumption Granger causes GDP of whether GDP Granger causes Electricity Consumption **Error! Reference source not found.**.

3.5 Data

The data used in this study was collected from the World Bank national accounts data, and OECD National Accounts data files for the Gross Domestic Product **Error! Reference source not found.**,The International Monetary Fund international financial statistics and data files**Error! Reference source not found.**, and the International Energy Agency data files on energy consumption**Error! Reference source not found.**. The selected range for the data is from the year 1985 to the year 2014. Fig.15. Electricity power consumption (kWh per capita), Fig 16. Consumer Price Index and Fig.17.GDP at current in $US illustrates the graphs for the electricity consumption from 1985 to 2014, the consumer price index from 1985 to 2014 and the Gross Domestic Product from 1985 to 2014 respectively. From Fig 16. Consumer Price Index, we note that electricity consumption has generally been declining since 1984 from just over 900KWh in the 1980s to somewhere below 600KWh just before 2010 and only saw a considerable rise between 2010 to 2014. The consumer price, on the other hand, has seen a general increase from 1985 to 2014 where it is noted to have increased from a rate from below 1 in 1985 to over 100 in post 2010. Finally, the Gross Domestic Product (GDP) has seen a steady increase from 1985 and saw a generally large increase from the mid 2000s to 2014. The data for electricity consumption however would, in the surface, seem to be paradoxical with that of population because under normal circumstances, because it may be assumed that electricity consumption increases with population.

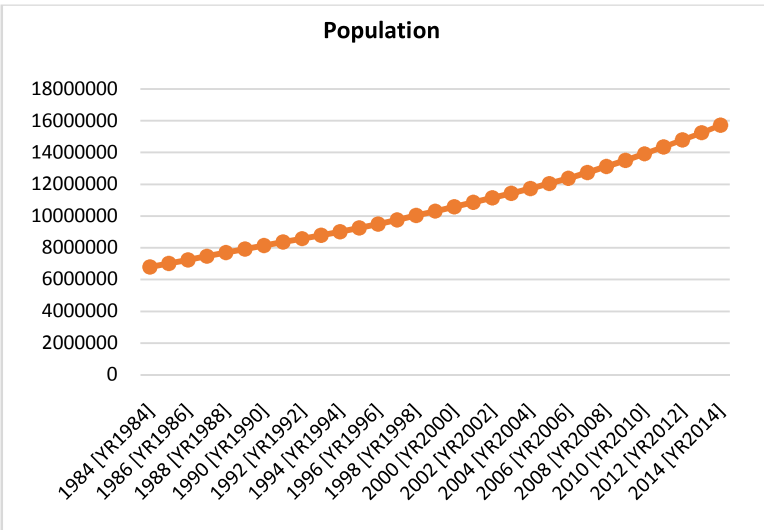

Fig.19. Population in Zambia

As seen from in Zambia, there is a general increase I population which may seem to be somewhat inverse to population growth however, there are other factors to consider such as the total population with access to electricity and other factors such as load shedding that have been conducted by the Zambia Electricity Supply Cooperation (ZESCO). The other variables on the other

hand seem to be positively related to one another.
The data analysis tool used is the R data analysis tool.
Table **5** shows the Descriptive statistics of included variables used in the study and Table 6, Table 7 and Table 8 shows the results of the Augmented Dickey-Fuller Test Unit Root Tests for the various variables.

Table 6: Augmented Dickey-Fuller Test Unit Root Test for consumer price index:

```
###############################################
#####
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################
#####

Test regression drift

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4416 -0.6763  0.1192  0.7953  1.9492

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.73658    0.42051   1.752   0.0921 .
z.lag.1      0.01465    0.01206   1.215   0.2359
z.diff.lag   0.77933    0.14454   5.392 1.36e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’
0.1 ‘ ’ 1

Residual standard error: 1.334 on 25 degrees of
freedom
Multiple R-squared:  0.8585,       Adjusted R-
squared:  0.8472
F-statistic: 75.84 on 2 and 25 DF,  p-value: 2.422e-
11

Value of test-statistic is: 1.2145 1.958

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.58 -2.93 -2.60
phi1  7.06  4.86  3.94
```

Table 7. Augmented Dickey-Fuller Test Unit Root Test for Electricity consumption

```
################################################
##
# Augmented Dickey-Fuller Test Unit Root Test #
```

```
###############################################
##

Test regression drift

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
     Min       1Q   Median       3Q      Max
-109.305  -14.166    5.051   16.191   93.854

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 128.45752   58.23689   2.206   0.0368 *
z.lag.1      -0.19497    0.08407  -2.319   0.0289 *
z.diff.lag    0.17548    0.17659   0.994   0.3299
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 41.72 on 25 degrees of freedom
Multiple R-squared:  0.2043,       Adjusted R-squared:  0.1407
F-statistic:  3.21 on 2 and 25 DF,  p-value: 0.05742

Value of test-statistic is: -2.3191 2.911

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.58 -2.93 -2.60
phi1  7.06  4.86  3.94
```

Table 8: Augmented Dickey-Fuller Test Unit Root Test for the gross domestic product at current

```
###############################################
# Augmented Dickey-Fuller Test Unit Root Test #
###############################################

Test regression drift

Call:
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:
       Min         1Q     Median         3Q        Max
-3.694e+09 -7.471e+08 -1.067e+08  7.105e+08
```

```
3.795e+09

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.476e+08  4.647e+08   0.748    0.461
z.lag.1      9.178e-02  5.188e-02   1.769    0.089 .
z.diff.lag  -2.283e-01  2.432e-01  -0.939    0.357
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.653e+09 on 25 degrees of freedom
Multiple R-squared:  0.1128,	Adjusted R-squared:  0.04182
F-statistic: 1.589 on 2 and 25 DF,  p-value: 0.224

Value of test-statistic is: 1.7693 4.536

Critical values for test statistics:
1pct  5pct 10pct
tau2 -3.58 -2.93 -2.60
phi1  7.06  4.86  3.94
```

Table 9: Descriptive statistics of included variables

Variable	**Definition**	**Usable observations (1985-2014)**	**Range**	**Mean**	**Standard deviation**	**Minimum**	**Maximum**
GDP	GDP at current in $US	30	29	8839092451	8420509841	1661948718	28045460442
CPI	Consumer Price Index	30	29	39.9	43.31508	0.01355	130.8
ELEC	Electric power consumption (kWh per capita)	30	29	695.2	102.7814	556.1	935.5

4. Results

As earlier stated, the Granger causality test was used in this study. The Granger causality test was performed in the R statistical package. For lag value of 1 was used in order to Granger test for all the hypotheses that were made. Table 10 shows the results for the for the Granger tests.

Table 10. Results of the Granger tests

Hypothesis	**P-Value**	**Result**
Rate of electricity consumption does not Granger cause the Gross Domestic Product (GDP)	0.5149	Cannot reject

Rate of electricity consumption does Granger cause the Gross Domestic Product (GDP) of the country	0.5149	Reject
Rate of electricity consumption does not Granger cause the Consumer Price Index	0.001077	Reject
Rate of electricity consumption does Granger cause the Consumer Price Index	0.001077	Cannot reject
Gross Domestic Product (GDP) does not Granger cause the rate of electricity consumption of the country	0.599	Cannot reject
Gross Domestic Product (GDP) does Granger cause the rate of electricity consumption of the country	0.599	Reject
Consumer Price Index does not Granger cause the rate of electricity consumption	0.4572	Cannot reject
Consumer Price Index does Granger cause the rate of electricity consumption	0.4572	Reject

The first set of hypotheses involve the relationship between Electricity consumption and the Gross Domestic Produce at current prices in Billions of US dollars. The null hypothesis (H_0) in this case is that the rate of electricity consumption does not Granger cause the Gross Domestic Product (GDP) of the country. For this set of hypotheses, we got a p-value of 0.5149 therefore, we cannot reject the null hypothesis. Therefore, because the p-value is not significant, we have to conclude that Electricity consumption does not Granger cause the GDP.

The second set of hypotheses involve the relationship between electricity consumption and the Consumer Price Index (CPI). The null hypothesis (H_0) for this set of hypotheses is that the rate of electricity consumption does not Granger cause the Consumer Price Index. The p-value for this set of hypotheses is 0.001077 therefore, we have to reject the null hypothesis. Therefore, because the p-value is significant, we have to conclude that the rate of electricity consumption does Granger cause the consumer price index.

The third set of hypotheses involve the relationship between Gross Domestic Produce at current prices in Billions of US dollars and the Electricity consumption. The null hypothesis (H_0) in this case is that the Gross Domestic Product (GDP) does not Granger cause the rate of electricity consumption of the country. The p-value for this set of hypotheses is 0.599 therefore, we cannot reject the null hypothesis. Therefore, because the p-value is not significant, we have to conclude that GDP does not Granger cause electricity consumption.

The final set of hypotheses involve the relationship between Consumer Price Index (CPI) and the electricity consumption. The null hypothesis (H_0) for this set of hypotheses is that the Consumer Price Index does not Granger cause the rate of electricity consumption. The p-value for this set of hypotheses is 0.4572 therefore, we cannot reject the null hypothesis. Therefore, because the p-value is not significant, we have to conclude that CPI does not Granger cause electricity consumption.

5. Conclusion

This paper examined the causal relationship between the rate of electricity consumption and the economy of a country using Gross Domestic Product and the Consumer Price Index in Zambia over the period of 1985 to 2014. The electricity consumption, Gross Domestic Product and Consumer Price Index were obtained from the World Bank on countries.

The results from this study are that there is a unidirectional relationship between electricity consumption and the consumer price index where the rate of electricity consumption Granger causes the consumer price index. The study also showed that there is no causal relationship between electricity and GDP and that there was no causal relationship between electricity consumption and the Consumer Price Index.

These results imply that a lot of goods purchased by consumers in the country largely electrically dependant. Also, by virtue of the fact that electricity consumption does not impact GDP in any significant form, we can also imply that there is either GDP activities are not wholly dependant on electricity consumption of that there has been little development or regression in GDP activities which makes it largely unaffected by electricity consumption. From these observations, a recommendation can be made that there is a need for an improvement in the production levels of electricity since we have observed that lower electricity consumption caused a growth in the consumer price index. This would imply that should electricity supply, and consequently, consumption were to increase, more goods would be produced and as a result, with more goods on the market, there would be a lowering of the consumer price index. From this study, it is also recommended that the Zambian government provide incentives to encourage the manufacturing of goods and services such as tax exemptions for any enterprising companies and de-regulation to assist in making the business environment more conducive in the country.

One area where this study can be expanded upon is that the same techniques can be applied with other areas of the economic and social setup in Zambia and to try and identify what exactly causes the effect between electricity consumption and the social/economic factor. For instance, one possible area of expansion is to further delve into how dependant Zambia's economy is on electricity and also provide solutions on how Zambia can improve/industrialize its operations.

References

[1] P. Mozumder and A. Marathe, "Causality relationship between electricity consumption and GDP in Bangladesh," Energy Policy, vol. 35, pp. 395-402, 2006.

[2] F. Jamil and E. Ahmad, "The relationship between electricity consumption, electricity prices and GDP in Pakistan," Energy Policy, vol. 38, p. 6016–6025, 2010.

[3] R. o. Z. L. Shedding, "Engineering Institution of Zambia," September 2015. [Online]. Available: http://www.eiz.org.zm/wp-content/uploads/2015/10/151008_Report_On_ZESCO_Load_Shedding.pdf. [Accessed 5 May 2017].

[4] Z. D. Agency, "Energy Sector Profile," September 2014. [Online]. Available: http://www.zda.org.zm/?q=download/file/fid/55. [Accessed 8 May 2017].

[5] S.-T. Chen, H.-I. Kuo and C.-C. Chen, "The relationship between GDP and electricity consumption in 10 Asian countries," Energy Policy, vol. 35, pp. 2611-2621, 2007.

[6] H. H. Lean and R. Smyth, "Multivariate Granger causality between electricity generation, exports, prices and GDP in Malaysia," Energy, vol. 35, no. 9, pp. 3640-3648, 2010.

[7] C. B. Jumbe, "Cointegration and causality between electricity consumption and GDP: empirical evidence from Malawi," Energy Economics, vol. 26, p. 61–68, 2004.

[8] A. Shiu and P.-L. Lam, "Electricity consumption and economic growth in China," Energy Policy , vol. 32, p. 47–54, 2004 .

[9] P. K. Narayan and A. Prasad, "Electricity consumption–real GDP causality nexus: Evidence from a bootstrapped causality test for 30

OECD countries," Energy Policy, vol. 36, p. 910–918, 2008.

[10] Faisal, T. Tursoy and N. G. Resatoglu, "Energy Consumption, electricity, and GDP Causality; The Case of Russia, 1990-2011," Procedia Economics and Finance, vol. 39, p. 653 – 659, 2016.

[11] Z. M. Fathimath and A. Geetha, "Analysis of Data Mining Techniques and its Applications," International Journal of Computer Applications, vol. 140, no. 3, pp. 6-14, 2016.

[12] T. W. Bank, "Electric power consumption (kWh per capita)," The World Bank, 2017. [Online]. Available: http://data.worldbank.org/indicator/EG.USE.ELEC.KH.PC?locations=ZM. [Accessed 2017].

[13] T. W. Bank, "Consumer price index (2010 = 100) | Data," The World Bank, 2017. [Online]. Available: http://data.worldbank.org/indicator/FP.CPI.TOTL?locations=ZM. [Accessed 2017].

[14] T. W. Bank, "GDP (current US$) | Data," The World Bank, 2017. [Online]. Available: http://data.worldbank.org/indicator/NY.GDP.MKTP.CD?locations=ZM. [Accessed 2017].