Collecting and Analyzing Energy Data for Central America, Mexico and the World: A Data Science Project

Leonel Morales, Krista Aguilar, Juan Ponciano, María Rivas

Instituto de Ciencia y Tecnología para el Desarrollo – InCyTDe
Universidad Rafael Landívar
{leonel, krista, juan, mafer}@incytde.org

Abstract. Data Science is a developing computational field in Central America and Mexico. The high availability of data related to energy and economy through several international agencies provides an excellent opportunity to plan and execute research projects and test tools and techniques. Also, energy is an important subject of research from different perspectives: economic, social, human development, engineering, among others. We took the challenge of locating the sources, collecting the data, cleaning it, performing initial exploratory data analysis, and applying basic modeling techniques to get familiar with the data and understand the information and knowledge it provides in regard to the possible nexus between energy and economic growth. In this paper we present a first recount of the process and some of the insights obtained during its execution.

Keywords: Data Science, Energy, Mexico, Central America.

1 Introduction

Nowadays data is being captured in multitude of environments, through sensors, logs, user input and other forms. The format in which data is being collected differs according to the needs of each project including plain text, numeric, binary formats, image, video, and application specific file formats. In general, it is a good idea to assume that there is no standard format. The extended use of computers, smart phones, tablets, and other devices with information processing capabilities increases the rate at which data is added to databases both in industrial and end-user settings. These data explosion phenomenon has been resumed with three words: volume, velocity and variety [3, 17, 25] although the discussion keeps going on about the convenience of including other words starting with V: veracity, volatility, validity, value, and others [11, 18, 24].

Four skills have been identified as relevant in order to get the most out of big data: programming, statistics, visualization and domain specific substantive expertise [10, 33]. Programming, or better, hacking skills are needed to download, parse, reformat, clean, and store data from several sources, statistics to explore and model relations in data, visualization to communicate findings visually through graphs and compositions

and domain specific substantive expertise to make sense of the whole process as it is always referred to a context which has its own rules and limits.

Although data science is not exactly the same as big data, a reference to it is needed in any data science project because it can be argued that data scientists need to be able to work with big data and in the process contribute the analysis, experimental design, and systematization that characterizes scientists when producing scientific knowledge. Several authors have abounded in this relation [5, 19, 25, 33]. For this paper the explanation provided in the previous paragraphs is adequate.

Our project concentrated in collecting data about energy and economy from several sources. Energy data and economy indicators are abundant (volume) come in different formats (variety) and are produced constantly (velocity). Programming skills together with statistical and visualization-building abilities come handy for getting and working with this data, which in turn is not possible without domain specific knowledge.

In order to be able to apply data science methods in this setting we started exploring data for Central American countries and then broadened our scope to include several countries from around the world. This exercise helped us to get familiar with the data and start identifying relations that are worth modeling in order to, in future efforts, describe the energy and economy phenomenon for our countries, do some classification and even predict potential outcomes. We consider that this effort will contribute to integrate energy data for the whole Central American region, help spread information useful for decision making, and also serve as input for further studies.

The following sections describe our data sources and the process to obtain the data (section 2), the exploratory data analysis up to this point along with a discussion of the insights obtained in it (section 3), and finally our plans to continue the project and the prospect models we intend to construct (section 4).

2 Energy and Economy Data Sources

Following the principles of the Open Data movement [26, 27, 28, 35], many institutions that collect information around the world have made it available through their websites opening new opportunities for researchers interested in applying data science methods and tools to contribute to obtain the knowledge hidden in it [8].

This has been the case in many fields like environment and ecology [31], biology [19], health [4, 16], finance and business [34], energy [15], and economy [2]. Governments have also agreed to make the data they collect available under certain guidelines to avoid compromising public security or exposing private personal information of citizens [6, 7, 12, 32, 36].

Not all data sources provide the same level of openness, or the same set of tools to work. Inconsistencies are not uncommon among different sources, presumable due to differences in methods, interpretations, conversions and integration criteria.

2.1 Selection Process

For this project a survey was made of the sites sharing information on energy and economy [21] see Table 1. There are an important number of institutions that provide that type of data, with different aims, scopes, and timeframes. Some of them are limited to a region, a country, or other geographical circumscription; others focus on renewable, non renewable or other subsets of primary sources. There are official agencies and NGOs dedicated to this task that share more or less information with different periodicity.

Based on consistency of publication and the broader scope possible, institutions and websites were selected to work with them. In the cases of energy data and economic indicators we found that information is published more consistently and for longer periods when it is released yearly, although other time intervals are possible.

Databases that include most countries were preferred to those that are limited to a single country or region, because the intention was to broaden the scope of the study. International agencies were found to be the most convenient.

Table 1. A list of sources of energy and economy data. In addition, most countries have local agencies for economy, energy and statistics.

Site	URL	Method for obtaining data
UN data	http://data.un.org/	API
United Nations Development Programme	http://hdr.undp.org/en/data	API
International Energy Agency	http://www.iea.org/statistics/	On screen tables
The World Bank, World DataBank	http://databank.worldbank.org/data/databases.aspx	API
Inter-American Development Bank, Statistics and Databases	http://www.iadb.org/en/research-and-data/statistics-and-databases,3161.html	Download as Excel file type
The U.S. Government's Open Data	http://www.data.gov/	API
Energy Statistics European Commission	http://ec.europa.eu/energy/observatory/statistics/statistics_en.htm	Download as PDF file type
Organization of the Petroleum Exporting Countries	http://www.opec.org/opec_web/en/publications/202.htm	Download as PDF file type
CEPALSTAT – CEPAL	http://estadisticas.cepal.org/cepalstat/WEB_CEPALSTAT/Portada.asp	API

2.2 Obtaining the Data

Although some sites allow downloading the data in a CSV or Microsoft Excel format, it is also common to find application programming interfaces – APIs – or web services that facilitate the process through programming.

In this project we also implemented a two phase method. In the first phase a Visual Basic for Applications – VBA – script inside a macro enabled Excel workbook was

crafted to open an Internet Explorer object that then navigated to a web page with the data of interest and transferred the data table into a new Excel workbook. With this method it was possible to set a list of web sites and a list of parameters to pass them to the script and let it download and save the data for later process. For certain websites, especially those designed to show the information online rather than to allow downloading it, this proved to be best suited.

In a second stage a Microsoft Access database was built and refined whenever needed, to grab and store all the information in the Excel workbooks. Again, a VBA script was crafted to automatically open each workbook as an Excel object, access each worksheet and insert the information into the tables of the database using SQL insertion queries.

This two phase method proved to be very useful. It came very handy to just enter the URLs in the macro workbook and then letting the script run frequently overnight to grab the data.

In some cases only the second phase was applicable, for example when the data was already available in an Excel file, which was the case for percentage of electric grid coverage by country that was located after some effort in that format. In other cases it was necessary to enter the information manually, locating it first in the web by searching and then entering it in the appropriate field in the database.

At this point it has been possible to compile information for more than 120 countries from around the world, for a period of time starting in year 1990 and ending in 2011, with most of the data covering intervals of one year and including data about energy and economy.

The database has more than 1.2 million registers and keeps growing.

3 Exploring Energy and Economy Data

Obtaining the data of interest, although very important and work-intensive, was just a first step in the data science project. The next, or better, the concurrent step as the first never ends, was getting familiar with the data, its meaning in the specific domain of interest – energy and economy – the way it vary among countries or intervals of time through the application of basic statistics, and showing it in graphics and charts to have a visual clue of where important phenomena may be occurring.

Crafting SQL queries was necessary to filter and extract the relevant data sets for the countries of interest that were later plotted. More than 80 SQL queries were programmed to produce 70 tables of data. 35 of them corresponded to Central American countries and 35 were the equivalent for a set of 11 countries outside of the region (Germany, Argentina, Brazil, Canada, Chile, China, USA, Spain, Japan, Mexico, and Venezuela).

Plotting compound indicators, those that result from ratios or percentages like electricity consumption divided by population, proved more convenient than quantifications like Gross Domestic Product (GDP) or Total Primary Energy Supply (TPES) for direct comparison among countries. On the other hand, the quantifiable energy and economy variables provide a contrast of size and volume so they are also worth plotting.

Some of the tables to chart were readily available from the collected data; others needed calculation from values taken from different places. Sometimes after calculating and plotting a table the resulting image signaled interesting relations or trends, for others the lack of relevant information was pointed to and so was the need to find and collect that information, and for most of the rest no interesting treat was found, although, arguable, the very exercise yielded knowledge about the data that could eventually be used in other context.

3.1 Plotting Data

To plot data in Excel in this initial stage, only one type of chart was used: lines with the indicator, value or ratio in the y-axis and the year from 1990 to 2011 in the x-axis. Although this can be considered only a very basic form of visualization, it helped to understand the evolution of each figure over time and the differences among countries. A brief discussion of a few of the charts analyzed in the project will help to appreciate the extent this effort may have with further refinements.

Electricity Consumption / Population. This ratio shows the gross amount of electricity consumed by the country (gross production plus imports less exports less losses) divided by the population. Fig. 1 shows the evolution over time for this indicator for Central American countries and Mexico. There are clearly two groups of countries according to how much electricity their inhabitants dispose of.

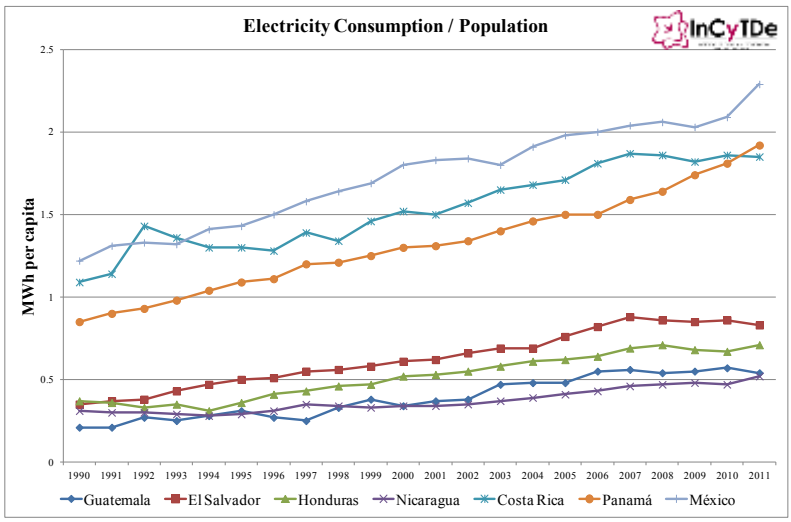

Figure 1. Electricity Consumption / Population. Mexico, Costa Rica and Panama provide, in average, more electricity to their citizens.

CO_2 / Population. This ratio calculates the average CO_2 emissions in tons per capita and is shown in Fig. 2. The amount of CO_2 can be linked to the sources the country employs for producing electricity and for other energy uses. A higher amount of oil and oil products may increase this ratio.

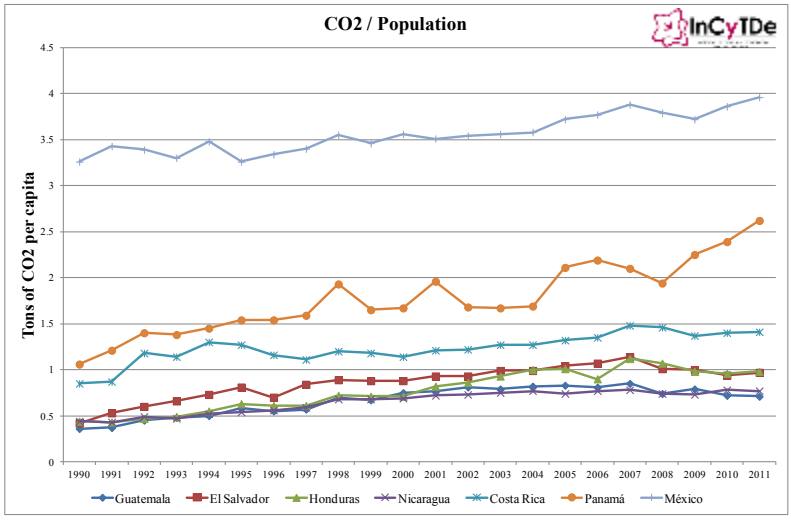

Figure 2. CO_2 / Population. The average of tons of CO_2 emitted per capita.

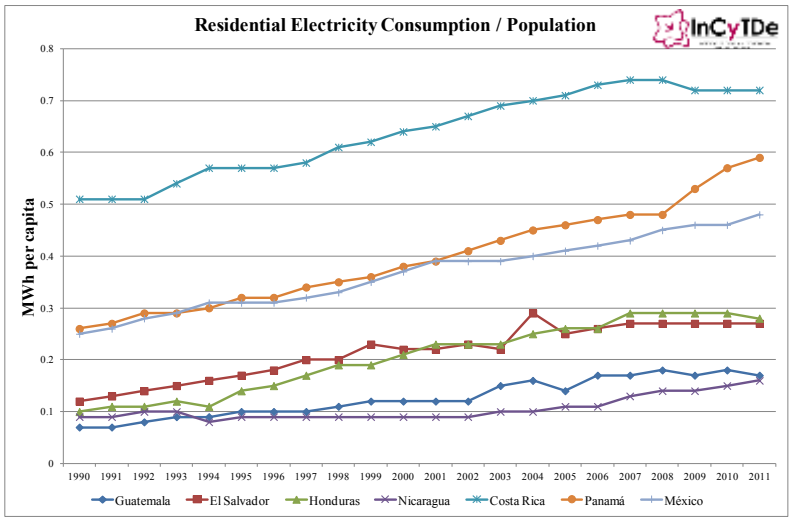

Figure 3. Residential Electricity Consumption / Population. This chart differs from Fig. 1 in that in this only the electricity used for homes is aggregated and averaged.

Residential Electricity / Population. Not all electricity consumed in a country is destined to housing needs; an important portion is used in commercial, industrial, services and others. This means that a more accurate figure for the availability of electricity in homes would be the amount of residential electricity consumed divided by the population. This is shown in Fig. 3.

Total Primary Energy Supply (TPES) / Population. TPES measures the total amount of energy available from primary sources before any transformation. Divided by population provides an indicator that is comparable among countries, see Fig. 4. This figure is indicative of how much energy the country disposes of. Transformations induce losses and inefficiencies so this indicator, in conjunction with others, may help identify them.

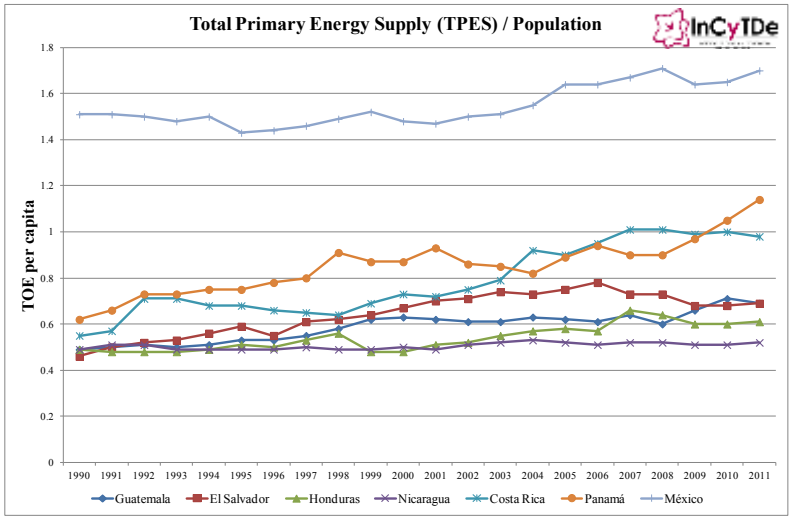

Figure 4. TPES / Population. In tons of oil equivalent or toe. 1 toe is approximately 11.63 MWh or 41.87 gigajoules.

Total Primary Energy Supply (TPES) / GDP. This ratio is key for the identification of the relation between energy and economic growth. As a country increases economic activities more energy is required. Although other energy variables can be used (energy imports, electricity consumption in the industrial sector, oil products consumption, and others) any relation can be referred to that of TPES and GDP. The indicator evolution over the years can be seen in Fig. 5.

Energy from Hydro / TPES in Balance. The Energy Balance Sheet keeps an account of the flows of energy, including transformations and final consumption, in a country [14, 37]. This chart is shows a trend in regard of energy from hydro sources, as shown in Fig. 6. As countries have increased the amount of energy they consume over the years they have tended to depend heavier on sources different from hydro.

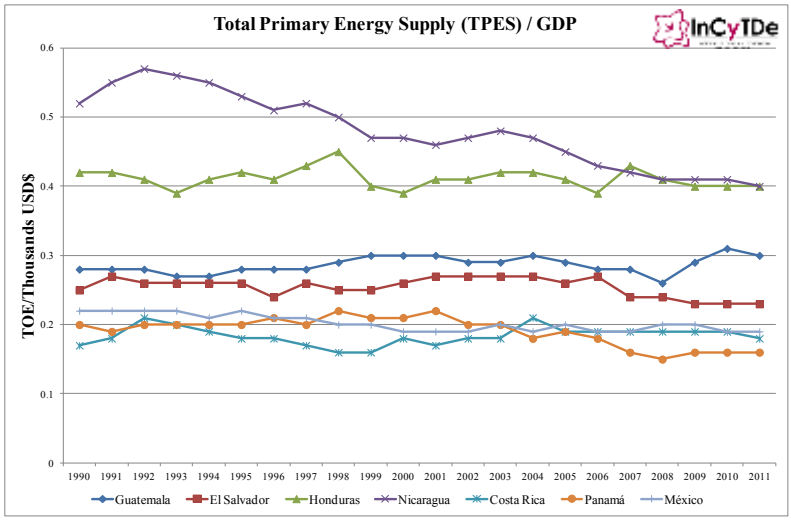

Figure 5. TPES / GDP. This chart depicts the close relation between energy and economy. A perfect relation would render a constant ratio. Except for the case of Nicaragua all the countries in the region keep a nearly flat curve.

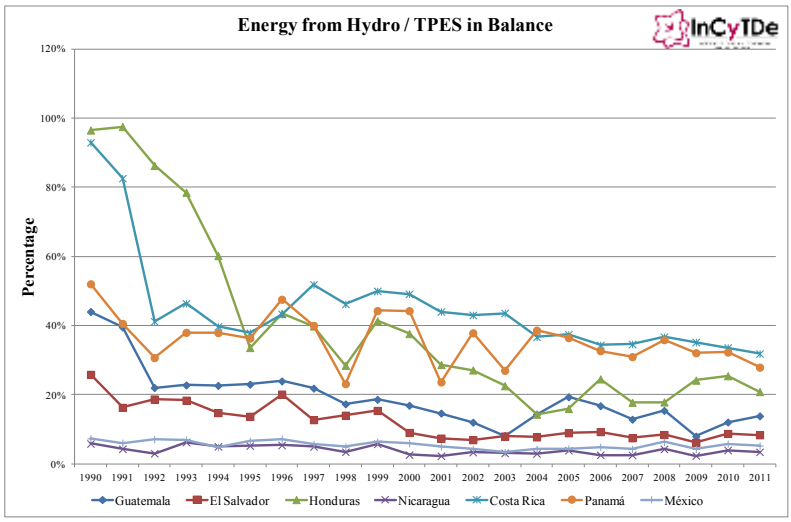

Figure 6. Energy from Hydro as a percentage of TPES according to figures in Energy Balance. Countries in the region have reduced the percentage of energy they get from hydro sources, not by reducing the net volume of production but by increasing other sources.

3.2 Beyond Charts

More advanced techniques of visualizations can be employed to analyze data. In a previous report [22] the use of Sankey diagrams or Energy Flow Diagrams was presented as a better option for comprehending the complete energy situation of a country. At this point in the project, all the information is available to produce such diagrams not only for a single year or a single country, but for several years and several countries.

This is a challenging task not only to craft the visualizations but to be able to compare among the different sets of data and find the trends and models needed. Nevertheless, we consider that building these visualizations is an important step towards a full understanding of the dynamics of energy and economy data.

For an example of the Energy Flow Diagrams see [13, 22] and Fig. 7.

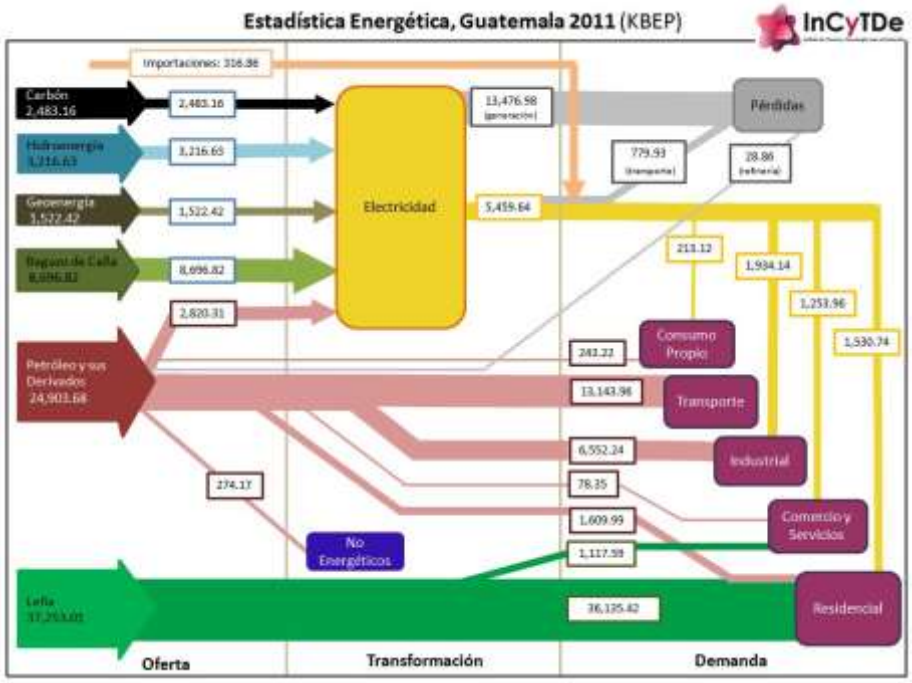

Figure 7. Energy Flow Diagram prepared in Spanish for energy data of Guatemala for 2011.

4 Proposing Models

Causal relationship between energy usage and economic growth is a topic under continuous examination in every country and proves to be central for discussions of energy policymakers. The evidence for an underlying relationship is typically obtained within bivariate modeling frameworks from energy data in close relation to

the main economic activities. As such, clean data about the composition of energy production and consumption, or the generation and consumption of energy by energy source and by economic sector constitute an essential input to those analyses.

Two interesting examples of analysis and modeling of energy and economy data, in line with this can be found in [1] and [29].

The common and simplest approach is to employ bivariate models, from which it is possible to capture the general ideas of the probable scenarios. However, many studies have shown that this first approach might lead to unsatisfactory or conflicting results, proving thus necessary to resort to more elaborated models [9].

There exist several efforts to build more specific models including other variables such as indicators in regard to capital measures and labor of force [23, 30, 38].

One promising line seems to be studying and modeling the behavior of aggregate energy variables like electricity consumption, total primary energy supply (TPES) and variables from the economy domain like GDP, population, inflation and others.

As Fig. 5 suggest, there is a good correlation between TPES and GDP for most countries of Central America and Mexico. It is easy to see that a clear next step is to find how good is that correlation for the rest of countries of the world and what are the characteristics of those with good correlation versus those with a less significant figure. Intuition suggests that countries with similar correlations between those variables, or any other more relevant pair, may present similarities worth studying.

Because energy consumption patterns vary in countries according to, among many other reasons, their geographical position, the integration of geo-positional information into the database seems to be relevant. We are incorporating such information now for future use. Under this perspective it can be argued that countries in similar latitudes should have similar energy consumption needs and that should be reflected in the data when isolated from the influence of other variables. This though, remains unverified.

Global energy data is difficult to produce and collect. Most of the databases consulted for this work have a delay of at least two years in their data, meaning that we will be able to see the fruits of current efforts in the field of renewable sources or changes in policies and production models until two years pass. Nevertheless as technology advances the recollection process will most likely speed up.

It is very important then to prepare the data science tools, including databases, data gathering processes, analytics, models and predictions for this speed increment.

5 Conclusion

In this paper we have provided a first account of the process we are conducting to apply data science techniques and methods to energy and economy data for the countries in the Central American region with the intention of applying the same analysis to the whole world.

We started explaining why this can be considered a data science project and how energy and economy data have the characteristics of big data. The basic sources to obtain the data were listed with notes about the method that can be employed with

each one, although the list is not exhaustive, it is a starting point for anyone wanting to apply data science in a practical manner.

How to get familiar with the data through plots, chart and diagrams was then explained and a discussion of the possible models was included. All these models and many others, limited only by the creativity of each data scientist, can be built, tested and applied in order to produce predictions that will be validated in the near future.

It can be argued that the main point of this whole article is to signal the opportunity energy and economy data presents to develop the data science computational field in Central America and Mexico but it will not be the main advantage for sure. We will all benefit from a better understanding of the dynamics of energy and economy in the first place.

References

1. Apergis, N., Payne, J.: Energy Consumption growth in South America: Evidence from a panel error correction model, Energy Economics 3, 1421-1426 (2010)
2. Beck, T., Demirgüç-Kunt, A., Levine, R.: Financial institutions and markets across countries and over time: Data and analysis. World Bank. (2009)
3. Birke, R., Björkqvist, M., Chen, L. Y., Smirni, E., Engbersen, T., Xue, J.: (Big) Data in a Virtualized World: Volume, Velocity, and Variety in Cloud Datacenters. In Proceedings of the 12th USENIX Conference on File and Storage Technologies (FAST 14) (pp. 177-189). USENIX (2014)
4. Boulton, G., Rawlins, M., Vallance, P., Walport, M.:Science as a public enterprise: the case for open data. The Lancet, 377(9778), 1633-1635 (2011)
5. Bryant, R., Katz, R. H., Lazowska, E. D.: Big-Data Computing: Creating Revolutionary Breakthroughs in Commerce, Science and Society. (2008) http://www.cra.org/ccc/files/docs/init/Big_Data.pdf
6. DATA.GOV.UK, http://data.gov.uk/
7. DATA.GOV: Frequently Asked Questions, http://www.data.gov/faq
8. Davies, T., Perini, F., Alonso, J.: Researching the emerging impacts of open data. World Wide Web Foundation. (2013)
9. Davis, C.: Making Sense of Open Data: From Raw Data to Actionable Insight. Dissertation. Next Generation Infrastructures Foundation. (2012)
10. Graves, S.: Meeting the challenges of data-intensive science. In Proceedings of the 2011 workshop on Climate knowledge discovery (pp. 4-4) ACM (2011)
11. Grimes, S.: Big Data: Avoid 'Wanna V' Confusion. InformationWeek.com (2013)
12. Huijboom, N., Van den Broek, T.: Open data: an international comparison of strategies. European journal of ePractice, 12(1), pp. 1-13 (2011)
13. InCyTDe: Balance Energético Guatemala 2011. http://incytde.org/flujo_energia/principal.html
14. International Energy Agency: Energy Balances Statistics. http://www.iea.org/statistics/topics/energybalances/
15. Krioukov, A., Goebel, C., Alspaugh, S., Chen, Y., Culler, D. E., Katz, R. H.: Integrating Renewable Energy Using Data Analytics Systems: Challenges and Opportunities. IEEE Data Eng. Bull., 34(1), 3-11 (2011)
16. Krumholz, H. M.: Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System. Health Affairs, 33(7), 1163-1170 (2014)
17. Laney, D.: 3D data management: Controlling data volume, velocity and variety. Technical Report, META Group Research (2001)

18. Laney, D.: Deja VVVu: Others Claiming Gartner's Construct for Big Data. In Gartner Blog, Jan.14 (2012) Available: http://blogs.gartner.com/doug-laney/deja-vvvue-others-claiming-gartners-volume-velocity-variety-construct-for-big-data/
19. Marx, V.: Biology: The big challenges of big data. Nature, 498(7453), 255-260 (2013)
20. McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., Barton, D.: Big Data. The management revolution. Harvard Bus Rev, 90(10), 61-67 (2012)
21. Morales, L.: Bases de Datos en Internet. In InCyTDe Blog, Apr.1 (2014) Available: http://incytde.org/incytde/content/bases-de-datos-en-internet
22. Morales, L. V., Aguilar, K. I., & Ponciano, J. A. Visualizing Energy Data and Seeing the Whole Picture of Energy in Guatemala. In Proceeding of Power and Energy 2013, ActaPress, (2013)
23. Nasreen, S., Anwar, S.: Causal relationship between trade openness, economic growth and energy consumption: A panel data analysis of Asian countries. Energy Policy, 69, 82-91 (2014)
24. Normandeau, K.: Beyond Volume, Variety and Velocity is the Issue of Big Data Veracity. Inside Big Data. (2013)
25. Novikov, B., Vassilieva, N., Yarygina, A.: Querying big data. In Proceedings of the 13th International Conference on Computer Systems and Technologies (pp. 1-10). ACM (2012)
26. Open Data Commons, http://opendatacommons.org/
27. Open Definition, http://opendefinition.org/
28. Open Knowledge: What is Open? https://okfn.org/opendata/
29. Ozturk, I., Aslan A., Kalyoncu H.: Energy consumption and economic growth relationship: Evidence from panel data for low and middle income countries, Energy Policy 38, 4422-4428 (2010)
30. Ozturk, I.: A literature survey on energy-grow nexus, Energy Policy 38, 340-349 (2010)
31. Reichman, O. J., Jones, M. B., Schildhauer, M. P.: Challenges and opportunities of open data in ecology. Science(Washington), 331(6018), 703-705 (2011)
32. Santana, M. T., & da Silva Craveiro, G.: Challenges and requirements for the standardisation of open budgetary data in the Brazilian public administration. In GI-Jahrestagung (pp. 836-848) (2013)
33. Schutt, R., O'Neil, C.: Doing Data Science: Straight Talk from the Frontline. O'Reilly Media, Inc. (2013)
34. Streeter, L. A., Kraut, R. E., Lucas Jr, H. C., Caby, L.: How open data networks influence business performance and market structure. Communications of the ACM, 39(7), 62-73 (1996)
35. The Open Data Foundation, http://www.opendatafoundation.org/
36. U. S. Office of Management and Budget (OMB): Open Data Policy – Managing Information as an Asset, M-13-13, http://www.whitehouse.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf (2013)
37. Vargas, A.: National Energy Balances. In Natural Resources Forum (Vol. 6, No. 1, pp. 29-42). Blackwell Publishing Ltd. (1982)
38. Yıldırım, E., Sukruoglu, D., Aslan, A.: Energy consumption and economic growth in the next 11 countries: The bootstrapped autoregressive metric causality approach. Energy Economics, 44, 14-21 (2014)