Analysis of "Greenhouse gas emissions by region" data based on CRISP-DM methodology

Assem Kussainova
Department of Computer Science
Nazarbayev University
Nur-Sultan, Kazakhstan
assem.kussainova@nu.edu.kz

***Abstract*—Greenhouse gases play an important role in the survival of humans and other living things by trapping some of the sun's heat and making our planet habitable, but industrialization, as well as deforestation and certain agricultural practices have led to an increase in emissions greenhouse gases. This paper provides a description of a greenhouse gas dataset based on the second phase of the CRISP-DM (Cross-Industry Standard Process for Data Mining) methodology Data Understanding. In this case, the emissions data were taken from a web portal from Stats NZ (New Zealand's official data agency) and then analyzed using the Microsoft Excel software. Using the analysis, you can suggest industries and regions in New Zealand where carbon dioxide emissions can be reduced.**

Keywords—CRISP-DM, greenhouse gas emissions, data analysis

I. Introduction

The importance of the data is growing synchronously with the development of the IT technologies. Recently, in many areas, the utilization of the Big Data has been increasing steadily.

In this paper, greenhouse gases emission data from New Zeland is described, based on the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology. This paper is organized as follows. Firstly, the theoretical background, including the definition of CRISP-DM methodology is overviewed. Then description of dataset is provided according to Data Understanding phase of CRISP-DM, concerning categories of data, data quality, discovered problems, and etc. Next, paper provides a conclusion of the description and recommendations what to do with the data. Finally, resources that were useful for the paper are listed.

II. Theoretical Background

Data Mining discovers statistically meaningful rules and techniques that automatically find patterns in large amounts of data [1]. There are several ways of performing Data Mining. In this paper, we follow the most commonly used CRISP-DM methodology. CRISP-DM methodology is a standard methodology for performing standard Data Mining process, which does not depend on a particular industry or a particular tool. There are various techniques of Data Mining, such as decision trees, clustering, regression analysis, and neural network analysis [2].

Data Understanding phase of CRISP-DM involves the collection of data and familiarity with information, identification of problems with data quality (errors or omissions) [3]. It is needed to understand what information is available, try to find interesting datasets, or form hypotheses about the presence of hidden patterns in them. There is no modeling at this step, only descriptive analytics are used.

III. Dataset Analysis

This phase includes 4 stages: Data collection, Data Description, Data Exploration and Data Quality. Following is the analysis of the dataset according to each of the stages.

A. Data Collection

The data about greenhouse gas emissions by industries and households used in this paper belongs to New Zealand's official data agency. This release of greenhouse gases by region (industry and household) includes data for 15 regions by main industries and households, so gas emissions information data is collected for 11 years from 2007 to 2018.

The estimates are compiled on the basis of the number of emissions on a production basis by economic residents. Region-level estimates are compiled using a top-down approach in order to maintain national consistency [4]. These are estimated at a fine degree of industry detail and aggregated for confidentiality purposes. However, several data sources used are based on unit record data. The dataset emphasizes that greenhouse gas may be emitted from multiple processes and covers emissions from New Zealand's agriculture, energy, IPPU, and waste sectors. Emissions are expressed in relation to carbon dioxide values.

Data has been written into one .csv file:

- greenhouse-gas-emissions-by-region-industry-and-household-year-ended-2018-csv.

B. Data Description

At this stage, using the available data, it is necessary to describe the data in all sources (table, key, number of rows, number of columns, disk space).

Dataset can be freely analyzed by any tool as it has open access. The size of the data is 316Kb and it is represented in the format of table with the total number of records of 3061. It contains seven attributes which are:

- region: values are the names of 15 New Zealand regions in string format;
- anzsic_descriptor: the values for this attribute are the names of industries involved in production of carbon dioxide in separate and together as total for all industries in string format (*"Total"* value is used in combination with CO2 separate components);
- gas: contains the *"Carbon dioxide equivalents"* values for mixture of gases and the separate names of

gases for representing amount of shares in pollution in string format;

- units: all amount of gases is measured in kilotons, which value is written in string format as *"Kilotonnes"*;
- magnitude: all data is assumed in the magnitude of CO2, and written in the string format as *"Carbon dioxide eqivalents"*;
- year: the period of time indicated from 2007 to 2018 in integer number format;
- data_val; amount of greenhouse gas emissions in the format of float number format.

C. Data Exploration

During this phase, visualization is used to explore the data and to capture interesting trends, as well as a list of attributes that are potentially useful.

The most meaningful attributes for exploring current dataset trends are those which represent industry sectors, regions and amount of gas emissions, while the least important are the units and magnitudes as they are the same for the whole dataset. Following is main useful information gained from dataset.

The main information gained from this dataset is that the industries have higher amount of gas emissions rather than households. Following is the chart where can be seen comparison of industries and households by regions.

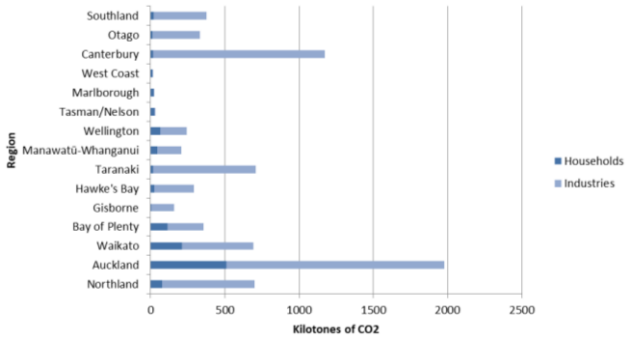

Fig. 1. Greenhouse emissions in agriculture industry during 2007-2018.

A special feature of New Zealand is the extremely high share of agriculture in greenhouse gas emissions. This sector, primarily livestock, accounts for about more than half of the country's greenhouse gas emissions.

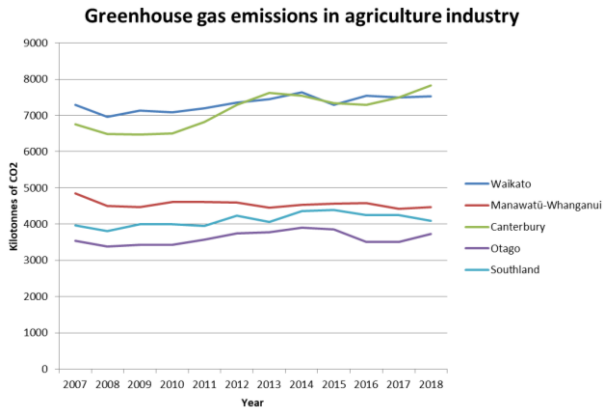

Fig. 2. Greenhouse emissions in agriculture industry during 2007-2018.

The highest rates of gas emission from agriculture industry are in the Waikato and Canterbury regions for 2018. Almost half less the amount of pollution occurs in the Manawatu-Whanganui, Southland and Otago regions, which emit 4500, 4000 and 3800 kilotons respectively.

Overall, this dataset can be further used for analysis of time series and prediction purposes, meaning how data changes over some time period for several variables with the help of Data Mining and Machine Learning techniques including Linear Regression, SVM, LSTM and etc.

D. Data Quality

This step requires an assessment of the quality of the data, as any inconsistencies can affect the progress of the project.

The dataset contains complete information without missing values and all necessary information for further data analysis is present, and moreover for better findings the most recent information of last years 2019-2020 can be added to the current dataset. However, the representation of data lacks integrity and not user friendly, so it should be restructured for more convenient usage and for decreasing data manipulation, as it takes time to preprocess data and understand the meaning of fields. Therefore, I would recommend working out the structure of the dataset, namely, get rid of duplicate values in attributes by creating additional columns where data on the total amount of emitted gases can be stored. In addition, it is proposed to create time period attributes that represent each year separately, thus reducing the number of records for each region of New Zealand. It is also helpful to remove the units and magnitude columns to reduce data processing, and instead add this information to the main dataset description.

In general, dataset follows the principles of accuracy, completeness, consistency, relevance and reliability, but there are problems with coherency, definition and time.

IV. Conclusion

In conclusion, this paper described the analysis of a dataset of greenhouse gas emissions in regions of New Zealand at the Data Understanding stage of the CRISP-DM methodology. The report describes methods for obtaining data, as well as the detailed structure of the dataset, including all attributes and their values. After examining the dataset data, following all the points, useful information was obtained and recommendations with improvements were given for further use in research.

References

[1] http://en.wikipedia.org/wiki/Data_mining.

[2] Yihua Zhang, Yuan Wang, Chunfang He and TingTing Yang. 2014. Research on Forecast Model and Application of Customer Loyalty under the Background of Big Data. International Journal of Multimedia and Ubiquitous Engineering. Vol. 9, No. 10, pp. 209-222.

[3] http://crisp-dm.eu/.

[4] Stats NZ. 2020. Approaches to measuring New Zealand's greenhouse gas emissions.