CAN GEOSPATIAL DATA TELL US WHERE TO PLACE MORE FUEL STATIONS IN HUNGARY?

Capstone Project Summary

Contents

CEU eTD Collection

Institute: Central European University

Name: BARNA DOMOKOS

Department of Economics – MSc Business Analytics

Date: 23 June 2020

Project Host Company: Starschema

Project Supervisor: Eszter Windhager - P. / Head of Data Science at Starschema

CEU CENTRAL EUROPEAN UNIVERSITY

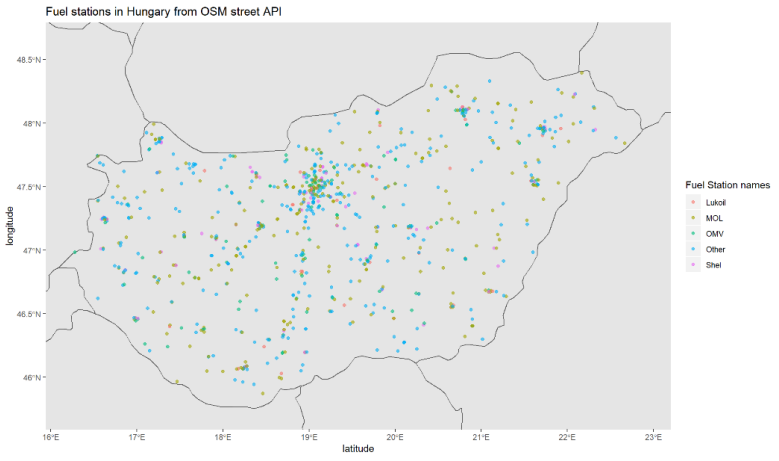
Section 1. About project host company and the project

Starschema is a Central European entity founded in 2006 where tech experts crunch and refine petabytes of data across every continent. Its profile covers data warehousing, business intelligence, big data services, and technology innovations for various large businesses around the globe. The Budapest-based tech company offers a large variety of services including data visualization and data science. They help to understand multinational companies how to analyse and efficiently use their data while creating trusted ground for business decisions such as business operation, investment and market expansion.

The business case that I had been provided to work on was chosen by the head of Data Science department of Starschema. One of their project focus is the analysis of the Hungarian market of fuel stations distribution, saturation of the fuel retail market, possible expansion for one of their large Client. The core challenge was to find new type of data such as geospatial data to identify if there are places where new fuel stations could be opened? Starschema guided me to work on one part of the aforementioned project. My assignment was to work on some parts of fuel market analysis. The direction and possible data source were named from where data can be freely accessed and downloaded. The focus point was the distribution of placed fuel stations in Hungary. I had to find ways to reveal fuel station market related insights, such as how traffic and related zone features relate to the presence of fuel stations.

Section 2. Market context*

Market share: more than 80% of the market is split among 5-6 players and respective distribution networks, franchise members. The major players are MOL, Shell, ÖMV (cca 70 % market share). The secondary players are Lukoil, Mobilpetrol, Auchan and others (cca 30% market share). The yearly Hungarian fuel sales is cca 3,6 billion litres (2019 data) and the estimated retail fuel sales market is worth HUF1370 billion (3.6 billion x avg price (HUF380)). There are around 2000 active fuel stations in Hungary. During this project, I have been working with around 800 fuel station data points that cover and represent more or less the full Hungarian market. For simplification, we assume that that data distribution is unbiased and samples the Hungarian fuel market. The plot below supports this assumption. **Avg turnover per shop:** (Y2019): HUF 7 billion. **Pricing strategy**: OMV, MOL, Shell – similar avg price level, Lukoil bellow the avg by 2-3 ft/litre Mobilpetrol: bellow the avg by 5-7 ft/liter. **Planning with Return On Investment time:** by domain experts individual fuel station ROI turning point on average were calibrated and expected to be on approximately 10 years period (calculated in books and business models). *Nowadays* strong expectation is tends to move up to a 5-year ROI turning point. The shorter time the higher is the pressure nn investors to make the right decision about the new fuel station placement. **How to identify white spot (high potential zone or point for placing a new fuel station) based on practice?** No formula exists for defining the potential of a certain location, experiences vary by market players and are subject to location type. The main drivers can be the traffic flow nearby and in the locations, regional economical development, competitors' presence.



* market-related numbers are estimated and indicative numbers that had been confirmed by a domain expert. Real market values might slightly wary from these figures.

Section 3. Direction for research and analysis

One contributing factor for running a fuel station profitable in the long term is positioning the price level of the sold products (fuel, food, and other items in the shop). Having no access to such public data other types of new data needed to be searched and explored. Public data that is available at the Hungarian Bureau of Statistics (KSH)

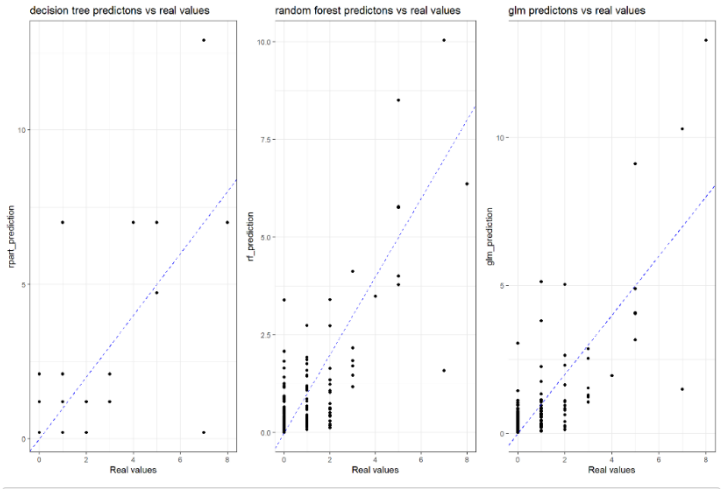
is measured traffic data (geospatial data, size of the traffic at a specific point collected in 2017). Questions I intended to explore were:

a) Are there places more fuel stations where the traffic is high and justifies the placement of a new fuel station?

b) Does traffic size explain a fuel station location placement or is it a reversed causality case?

c) From the available and free usability data what can be explored? Does the number of shops in the neighbourhood and traffic size explain the number of fuel stations? Building predictive models to reveal the possible relationship among several variables of interest might seem a good direction for research of this capstone project. I have chosen to work along this path.

Section 4. Proof of Concept



The core idea behind testing the concept can be briefed as: generate rectangles on the map of Hungary in such a way that each rectangle can be uniquely identified and their ID used by aligning function. Then we can align different geospatial data points to them (RHS variable data points). There are many options available to generate the map and map split using the geolocation coordinates (1km, 10km, 100 km size options are available). After matching each variable data range to corresponding rectangle ID (with its respective geocode data range behind) we have at hand the proper dataset with all important attributes to run the predictive models on them. Have chosen 3 models that seemed proper to run regression analysis on the data (tree, RandomForest, GLM with Poisson distribution, having at hand discrete integer numbers to work with and not continuous values of respective independent variables).

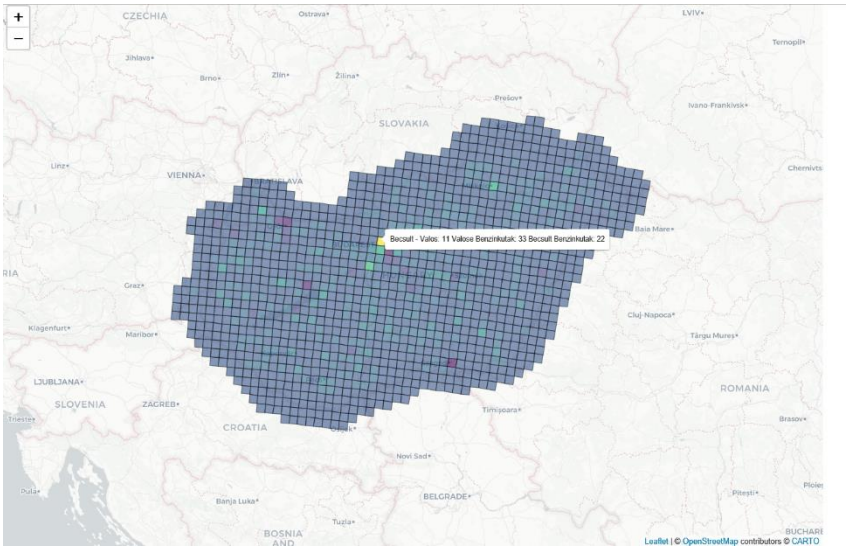
The applied 3 prediction model to compete simple Decision Tree, RandomForrest, GLM with Poisson distribution. The output the predictions shows that we can explain to a certain extent the number of fuel stations with RHS variables. There is such variation in the RHS variables (traffic-related aggregated features, number of shops, markets, restaurants (also the variable importance plot also confirmed this) - that it can be used to explain the variance of the LHS variable – number of fuel stations available. This is interpreted through the three output plot of predictor models. Out of the three models **RandomForrest proves to be the best predictor model producing the best RMSE value of 1.151061 versus the CART model value of 1.3327566 and the Generalized Linear Model of 2.278326.**

The core insight of all three model output is that there are cells where Prediction values versus Real values are higher. If there would not be variance all data would be compressed along the diagonal line – too ideal situation to be true. In other words, where the predicted value is higher than the real value there is the potential for investment in fuel station (the area can bear more station based on comparison with other similar composition cells).

CEU eTD Collection

Section 5. Prediction Results

My main goal was to identify cells or call them 10 x 10 km territory raster (the split size of a map raster applied within this capstone work) where investors should consider placing fuel station:



The mapping function split Hungary territory in a total of 1270 cells. Out of these total cells in some already do have fuel stations located and in some, there is no placement at the moment.

Some key figures of retrieved matrix information at the end of the prediction phase:

- Out of **1270 cells, 105 cells** can be considered for further analysis (**investment consideration potential** cells).
- Out of total potential cells, there are **69 cells** where can be placed **minimum** 1 fuel station.
- Out of total potential cells, there are **7 cells** in which there could be placed **more fuel stations next to the existing two** (for example a third one)
- There are **4 cells** where 2 more fuel stations can be placed next to existing ones.

Identifying where are these cells are located on the map can be retrieved and displayed by zooming in and hovering the mouse over the map and cells. It automatically displays on a label the cell information. Also, the colours of cells indicate the potential cells where the applied prediction model suggests fuel station placement.

I need to emphasize here again that predicting the revenue potential, profitability per shop was not subject to the study of this capstone project and in my opinion, would be worth investigating that part too within another research project.

Section 6. Conclusion and suggestion for further analysis

Car engines drives cars. Fuel consumption of car engines boosts or halts fuel retail markets, oil exploration and supply. Market demand for such utilities can be predicted quite well. Or not. See the turmoil in the demand market caused by COVID-19. Among such turbulent economical environment new tools methods are needed to bring up new insights about the continuously changing demand for utilities such as fuel and adjust investment decisions accordingly. It is true that change in a market structure can open up investment possibilities for investors. How to find them? Based on experts opinion is one way. One of the other ways is by analysing the right kind of data, looking on the market potential from different perspectives and base investment decisions on quality and relevant data. My attempt through this capstone work was to look on the market through available geospatial data points and run predictions using such category data. It proved to be a right approach.

Fresh insight came through near the project ending: we need to better understand what drives people, what triggers their behaviour, what incentivise them to jump in and use car instead of public transportation, walk or bicycle. We also need to better understand the motivation behind car ownership. There is new opportunity in studying behaviour patterns from digital traces one lives behind every day. Also with other type data that is generated on daily base (example WAZE app usage data traces) randomized controlled experiments (A/B tests) could be done without hustle and for example measure impacts of fuel or other merchandise pricing policy of the respective company on its revenue stream (one example: can we alter ones route by WAZE advertised incentive to buy fuel or merchandise in different location?). The obtained insights from such research could be utilised for example supporting or blocking a potential fuel station investment decision. Or at least putting more subtle questions on the table that needs to be answered before a fuel station investment gets green light.

CEU eTD Collection