Investigating Hungarian Agricultural Subsidies Data

Project Summary for CEU Business Analytics MSc

Gergo Szekely

2020 June

Contents

1 Client Introduction

The client of my capstone project was K-Monitor, an anti-corruption, watchdog organization. As their website states stakeholders of K-Monitor firmly believe that corruption is a problem that affects all parts of the society and harms everyone regardless of their beliefs, ethnicity, gender or political orientation. It not only causes enormous economic harm, but also undermines trust in institutions and the rule of law. It endangers democracies and also hurts people in regions where good governance would be must to eliminate hunger, poverty and violence. The best antidote against corruption is a society where citizen have a feeling of ownership over institutions and actively shape their environment. As a non-profit, K-Monitor supports institutions, journalists and individuals to fight corruption through community building, technology development, advocacy and research. Principles of their operation are openness, independence and a critical approach. I share these values and I would like to use my technical skills to help this never-ending fight against corruption.

2 Project Description

Hungarian Agricultural Subsidies are transferred to tens of thousands of beneficiaries each year but the majority of the funds enriches a small fraction of the beneficiaries. The money distributed to Hungarian farmers is 1.5-2% of the country's GDP so it is no surprise that many people want to get a share of it. Due to EU regulations detailed datasets have to be disclosed about the payments. These are published quarterly on the website of the Hungarian Treasury. Similarly to other publicly available datasets, it is not easy to process this information in order to understand and evaluate the system. K-Monitor has been working on this topic since 2014. By analyzing the data, one could find interesting patterns about how this money is spent. If it were available in an enriched, user-friendly format investigative journalists, non-profits or curious citizens could help the general public learn about this domain. The freudulent nature of this area was also highlighted by a NYTimes investigation in 2019.

3 Project Summary

This capstone was primarily a data engineering project but it touched almost all of the topics covered during my studies at the Business Analytics program at CEU.

3.1 Date Engineering

When creating the data engineering pipeline to perform error corrections and enhancements the aim was to make automatic. Anyone who has the code checked out from the project's public GitHub repository and has the runtime dependencies (Bash, KDB,R and Python should be able to build and run everything on a personal computer.

I downloaded the raw dataset from the Hungarian Treasury. Cleaning the addresses was the most complicated part of data processing. There are 3 fields that contain geographical information: zip code, settlement and address. During exploratory data analysis it became obvious that the address field was too varied to standardize by simple rules so I used the Google Maps Geocoding Api to convert the 426 000+ unique addresses. This reduced the number by more than 100 000 so aggregating became much more accurate.

I used an excel file from the Hungarian Central Statistics Agency to get how the country is divided into regional hierarchies. I joined this data to the agricultural subsidies so it is easier to extend to other statistical information.

I added gender information to individual beneficiaries by splitting citizens to the binary "male" and "female" based on the given names and common patterns. I could categorize most winners using the legally allowed name lists and manual overrides. This allowes analyzing gender-based distribution of the subsidies.

3.2 Findings

The clean dataset allowed high-level analysis and in-depth investigation targeting zip-codes, individuals or specific firms. When looking at the yearly distribution of subsidies summed by beneficiaries I saw a steady increase in the mean and median values. I saw that the number of payments almost doubled since 2010 but the number of beneficiaries went down. This means that fewer beneficiaries receive more money but in smaller installments. The top 10% accounts for more than 75% of the overall payments.

About 20% of the subsidies that are received by individuals goes to females while the rest ends up with men. The average amount of money won by females is 2/3 of what men receive.

When looking at the aggregate wins by institutions sharing the same address we see that the number of entities who received agricultural subsidies was higher in 2019 than in 2010. The mean increased by 10 mm HUF to just over 38 mm and the median more than doubled. The largest sum won by a single firm was almost 4.5 bn HUF in 2019.

The total amount of agricultural subsidies distributed in Hungary since 2010 is 6 307 bn HUF. From this amount 895 bn went to beneficiaries where both individuals and institutions share the same address (>14%).

Address conversion by the Google Maps API had another benefit besides standardizing the location information: latitude and longitude information is now available along with the original fields so I could add map-based visualizations.

I could also run simple aggregate queries that were impossible previously. I was interested to see if there are individuals who share the same address. There are a few addresses that would be interesting to investigate: some locations have dozens of distinct individuals winning hundreds of millions of HUF in subsidies. There might be a legitimate explanation behind this but it is certainly suspicious.

I received high-level financials of firms for 2018 and I joined these with the agricultural subsidy dataset. This allowed the cration of statistical models to find non-trivial patterns in the data. I defined two variables to measure how dependent a firm is on agricultural subsidies: a continuous and a categorical. The continuous variable is the ratio of agricultural subsidies won by the firm compared to its total sales. I also created a boolean variable from the "subsidy to sales" ratio by defining a threshold: if the total subsidies won by a firm

are at least 50% of its sales then it is dependent on this source of revenue. I created a few prediction and classification models to investigate these variables.

4 Next Steps

It would be impossible for a non-state actor to analyze the financials of individual beneficiaries because of the lack of public data. For firms commercial datasources exist but I only used a small fraction of the data. The external validity of the models I created could be improved because they do not say anything about individuals and even from the firm data a lot of observations were dropped either because of matching issues or missing values. Changes in political affiliation of an area and changes in regulations can change how firms treat agricultural subsidies. Trivial extensions would be settlement-based statistics from KSH, election data and more detailed financial metrics for firms.

The database I built can already be used for running complex queries but starting the service and constructing the queries requires technical skills. It is an obvious next step to create a simple GUI so journalists or other interested parties can do ad-hoc analysis easily. Since geocoordinates are now also part of the dataset a map-based tool can also be created.

5 Summary

Agricultural subsidies are the biggest chunk of the budget of the European Union. The system has many flaws but the stakeholders are not motivated to change the status quo. The problem domain could be better understood by analyzing the data. Targeted investigation for politically exposed people has been conducted but it is not possible to do even simple data exploration across multiple years. I did the first steps to solve this by taking the raw data and applying data engineering techniques. I only used free tools and made the code available on github. I did statistical analysis on the data and created dozens of visualizations. Even though the focus of the project was on the data engineering part I also built a few models using the cleaned agricultural subsidies and a small dataset about firm financials. It was a great technical challenge and I learned a lot while going through the whole data science process. I also learned a lot about agriculture and I will try to complete some of ideas in the `Next Steps` section in the next years.