Investigating Hungarian Agricultural Subsidies Data

CEU Business Analytics MSc

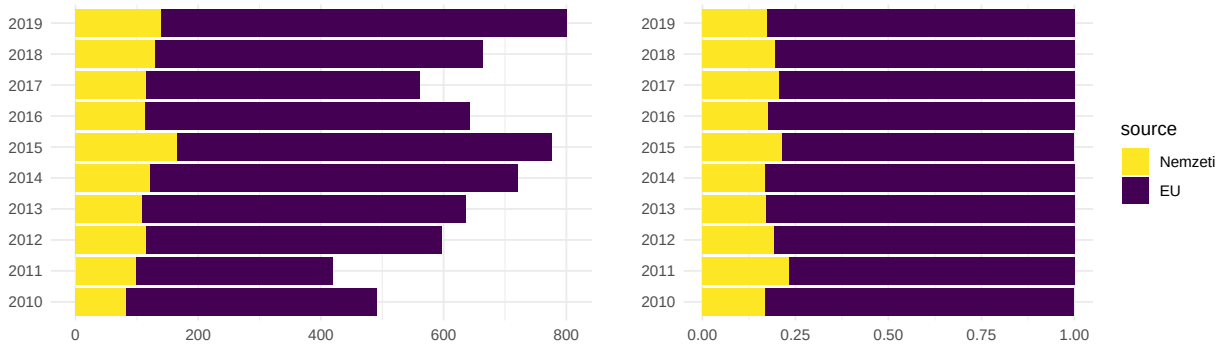
Gergo Szekely

2020 June

Contents

1 Problem Definition

Agricultural subsidies represent a huge section of the budget of the European Union and national governments. The way this money is distributed is often inefficient and sometimes fraudulent as covered thoroughly by the New York Times and various Hungarian news sites (here, here or here). The amount of 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. By analyzing the data we could find interesting patterns about how this money is distributed.



(a) Sum amounts in billion HUF

(b) Proportion of the groups

Figure 1: Yearly proportion of subsidies based on the source

The data is freely available on the website of the Hungarian Treasury but it is difficult to work with. Most of the money comes from the European Union so at least for the European Authorities it would be easy to spot if it was manipulated. Very limited analysis has been done on this data except for some high-level statistics and targeted investigations for high-profile actors; those concluded that the numbers were accurate.

Similarly to other datasets made public by government agencies, this one is difficult to work with for various reasons. To hide the true nature of their activities, state actors often comply with their own laws only to a minimal extent. They will make public whatever they are legally required to do so, but they are not motivated to make it easy to use. Another important reason for the badness of the data is that state agencies do not represent an attractive workplace for highly skilled workers. As a result, government agencies simply lack the talent to make great, public datasets. Cleaning and enriching the data would allow investigative journalists, non-profits or curious citizens to learn about this domain. This paper covers the initial efforts to solve this problem. The data transformation and analysis scripts are all publicly available on GitHub.

2 Data Processing Pipeline

The data has to go through a cleaning and transformation pipeline to perform various error corrections and enhancements. There are other, freely available datasets (e.g.: socio-economic, elections) that can be merged with the agricultural subsidies to provide a wide range of descriptive variables.

When creating the data engineering pipeline the aim was to make it as automatic as possible. Anyone who has the project checked out from the GitHub repository and has the runtime dependencies in place should be able to build the project from scratch.

2.1 Runtime Dependencies

Data processing and analysis can be performed on a personal computer using the following open-source or free tools.

- **Bash**: The various data sources can be downloaded and processed by executing a handful of shell scripts. Thus a unix environment with a command line is required.
- **KDB**: To efficiently process the raw data and merge it with other datasets I use a programming language called q that comes with an in-memory database solution called KDB. It is free to download for non-commercial use.
- **Python**: The project contains a script to call an external API to geocode addresses. This creates a standardized format for the location information and also adds latitude and longitude coordinates.
- **R**: Data analysis is done in R so a runtime is also needed.

2.2 Download Raw Data

The raw dataset is available from 2011 on the website of the Hungarian Treasury. The site provides an on-line search interface and links to compressed csv files that contain yearly batches of the subsidies. The website has direct links to the datasets from the past 5 years but based on the patterns it is possible to figure out how to download earlier batches and luckily most of the links still work. Data for 2010 used to be available on the site but it is no longer the case; I added that as part of the project and adjusted the download scripts to handle that in a transparent way.

The raw files are quite large:

```
## 39M ../input/csv/old_2010.csv    41M ../input/csv/utf8_2011.csv
## 40M ../input/csv/utf8_2012.csv   45M ../input/csv/utf8_2013.csv
## 49M ../input/csv/utf8_2014.csv   69M ../input/csv/utf8_2015.csv
## 69M ../input/csv/utf8_2016.csv   81M ../input/csv/utf8_2017.csv
## 82M ../input/csv/utf8_2018.csv   85M ../input/csv/utf8_2019.csv
```

2.3 Fix Character Encoding

The original files are in `ISO-8859-2` encoding (aka. `Latin-2`). After converting the data to `UTF-8` it is in a format that programming languages can deal with in a more natural way. Here are some sample rows:

```
## "Ábel Ferenc";6000;"Kecskemét";"Bódog sor 6";"Területalapú támogatás";"EMGA";"EU";81175
## "Ábel László";8330;"Sümeg";"Tik hegy 15";"Területalapú támogatás";"EMGA";"EU";1004025
## "Ábel Tibor";7743;"Romonya";"Petőfi utca 6";"Területalapú támogatás";"EMGA";"EU";397757
## "Abonyi Ferenc";4646;"Eperjeske";"Béke út 27.";"EMVA - Szakképzés";"EMVA";"EU";12868
## "Abonyi Ferenc";4646;"Eperjeske";"Béke út 27.";"EMVA - Szakképzés";"EMVA";"Nemzeti";4452
## "Abonyi Mónika";5081;"Szajol";"Major köz 1.";"Területalapú támogatás";"EMGA";"EU";392159
## "Ábrahám Dezső";8595;"Kup";"Malom utca 21.";"Területalapú támogatás";"EMGA";"EU";144156
## "Ábrahám Géza";8595;"Kup";"Fő utca 11.";"Területalapú támogatás";"EMGA";"EU";221132
## "Ábrahám Gyula";8919;"Kustánszeg";"Kossuth utca 74";"EMVA - Szakképzés";"EMVA";"EU";44576
## "Ábrahám István";6786;"Ruzsa";"tanya 471";"Területalapú támogatás";"EMGA";"EU";97690
```

2.4 Basic Data Cleaning

During exploratory data analysis I found multiple versions of some entities. Fixing this is important if we want to get an accurate view on the aggregate subsidies by beneficiary.

- **White spaces**: Some names contained multiple spaces when separating name parts. I replaced all whitespace group with a single space in the names.
- **Character cases**: Some names had multiple versions with inconsistent capitalizations (eg. all caps or 2 upper-case starting characters) so I split names by spaces and capitalized only the first character in each name part. Unfortunately, the built-in `upper/lower` functions of Q could not handle Hungarian characters with accents so I had to implement custom logic to do these transformations.
- **Remove dots**: Some names had multiple variations based on how abbreviations in their name parts were written: some had "." while other rows did not.

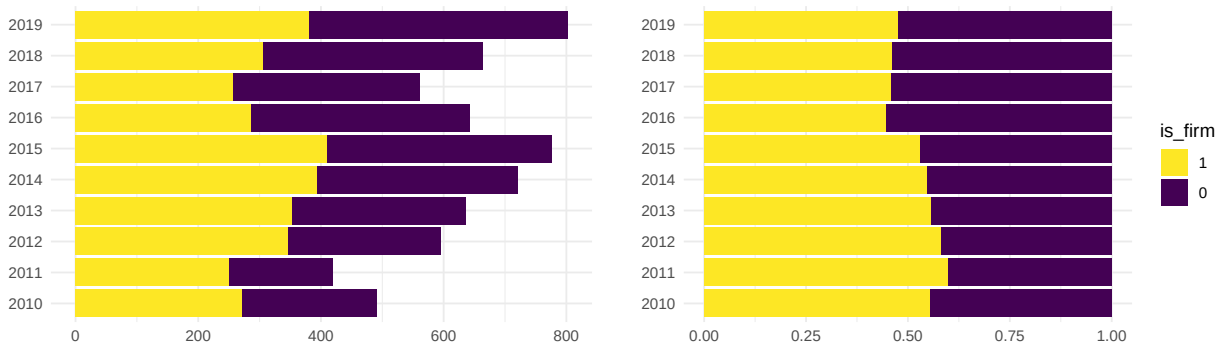
The above steps reduced the number of distinct beneficiaries by 82 023 from 532 221 to 450 198.

2.5 Identify Institutions

Individuals and institutions are not separated in the data. There is no indication of the beneficiary type but it would be important to investigate the 2 groups separately. I made the initial assumption that any name with 8 or more name parts is a firm. For shorted=r names I identified a large number of keywords that indicate that the beneficiary is an institution. Here are some of the keywords:

```
##  Megyei  ÁfÉsz Agrár kamara Agrárkamara Alapítvány Állami Apostoli Exarchátus Asztaltársas
## ág Baráti Kör Baráti köre barátok Kör Baromfi Batári Kör birtok BT Club Consulting Csoport
## osulás dísznövény Egyéni cég Egyesülés Egyesület Egyetem Egyház Egylet Együttműködési Erdő
## gazdálkodás Erdőszöv és vidéke Faiskola Fióktelepe GAMESZ Gazdakör Gazdaság Gondnokság Gyü
## lekezet Hegypásztor Kör Helytörténeti Húsüzem hivatal Igazgatóság Intézet Intézmény Iparte
## stület Iroda Iskola Kereskedelmi Kfc KFT Kht Kincstár kkt Klub Konzervipari Kórház Közössé
## g Központ Község Lelkészség llamkincst Lovas Kör Megyei Jogú Mesterséges Minisztérium Műve
```

As the above list shows there are not just firms but churches, municipalities, clubs and all sorts of other organizations; for simplicity I call all of these firms. Their distribution is around 50% for all years.



(a) Sum amounts in billion HUF

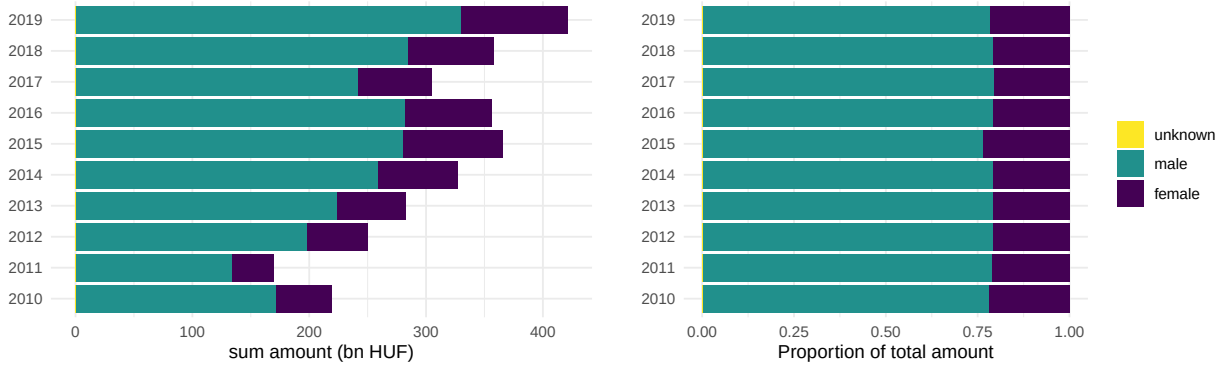
(b) Proportion of the groups

Figure 2: Yearly proportion of individuals and firms

2.6 Gender Information

Most of the non-institutional beneficiaries are Hungarian citizens. It would be interesting to see the gender-based distribution of the subsidies given to people. This information is not in the data but we can split citizens to binary "male" and "female" categories based on their names. Since there is no other information in the data this is the only way to split the people. All given names that are legally allowed in Hungary are

available on the website of the Hungarian Scientific Academy. Names not in the lists are either of foreign origin or contain some typos so they will be categorized as 'unknown'. Luckily, the amount of `unknown` records was not too much so I manually created additional lists to help categorize the unknowns. Some female citizens changed their legal names to that of their husbands by adding a "-né" postfix. This pattern is easy to find so that is also taken into account.



(a) Sum of subsidies won in billion HUF

(b) Proportion of the groups

Figure 3: Distribution of beneficiaries by gender

As the charts show about 20% of the subsidies that are received by individuals go to females while close to 80% end up with men. When summing the total wins by individuals the average amount of money won by females is consistently about 2/3 of what men receive.

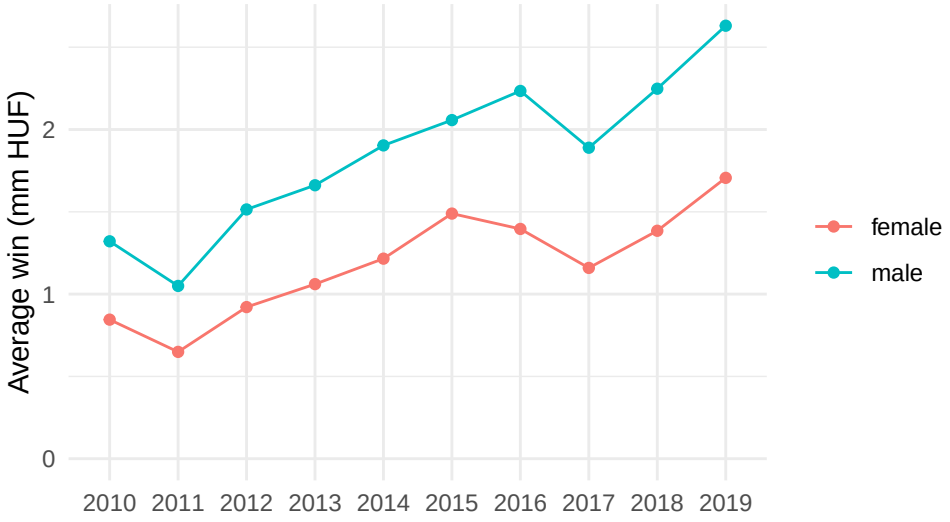


Figure 4: Average wins by gender

2.7 Clean Addresses

Fixing the addresses was the most complicated part of the data engineering pipeline. 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.

2.7.1 Missing addresses

The address fields are missing for almost 300 data points totaling almost 9.6 billion HUF. Some entries do not even have names but according to domain experts this is expected as these records indicate money that was used to cover the costs of running the program. This accounts for almost 9.3 billion of the incomplete records. There are also data points that do not have addresses but have names. If a name appears with exactly two variations: once with address and once without I assume that it is the same beneficiary and use the address. If there are no entries with the same name with filled addresses or there are multiple such entries in the full dataset then I cannot fix the address field and dropped these records. These amounts are tiny compared to the whole dataset so these should not have a real impact on the overall analysis.

2.7.2 Geocoding

To be able to aggregate agricultural subsidies on the beneficiary level it was important to have a geographical information that can be used as a key along with the name. After several iterations of unsuccessful address cleaning efforts I came up with the idea to use some geocoding API to convert all addresses to a standardized format. Unfortunately, I did not find an accurate, free service so I ended up using the Google Maps Geocoding API. The raw data has 426 675 unique addresses and each one of these had to be sent to the API for geocoding. I used a python script to create batches of addresses to geocode and process them. The service is free only up to a limit and geocoding all addresses would take almost a day without parallellization. I saved down the results so I wouldn't have to do this slow, somewhat manual process another time for addresses that were already geocoded successfully. The API is sensitive to permutations in the address fields so I tried to find an order that yielded the highest success rate. Unfortunately, Google Maps could not convert some addresses so in those cases I just use the original format. One extra benefit of this approach was that all addresses that could be processed now have geo-coordinates. Using the latitude and longitude information we can create a map showing in which areas those beneficiaries reside who got most of the subsidies. After geocoding the number of unique addresses decreased to 320 298 by 106 377 so aggregating will be much more accurate.

2.7.3 Geographical Information and Statistics

The Central Statistics Agency (Központi Statisztikai Hivatal - KSH) has a website where a lot of descriptive statistics are available about Hungary. A complicated excel file can be downloaded and parsed to get how the country is divided into regional hierarchies. KSH has their own ID to tag areas called `KSH_KOD`. For the majority of cases this ID represents a settlement and contains all the zip codes of that area. Unfortunately, after several attempts to join the 2 datasets it turned out that zip codes and `KSH_ID`s are in a many-to-many relationship. The data provided by KSH and the agricultural subsidies had reconciliation issues too. Some zip code + settlement pairs were different in the 2 datasets while other zip codes were missing altogether from the KSH tables but were valid (as verified by online map services). I reached out to KSH for clarification and they acknowledged the data issues. However, they will not take steps to fix them in this fundamental data source, rather redirected me at yet another dataset maintained by the Hungarian Postal Service that truly has all of the zip codes of the country. The iterative data cleaning and exploration process highlighted that there were still a few inaccuracies but those could be fixes by a small override file that is part of the project. Using these 2 datasources and a refined logic to fix mapping issues I was able to join the datasets.

2.8 Normalization

The full loaded and enhanced dataset is quite large:

```
## 1.0G ../output/agrar_full.csv
```

To eliminate redundancies and make the dataset more compact we can normalize it by extracting various aspects of the data like subsidy type, settlement information and the beneficiaries.

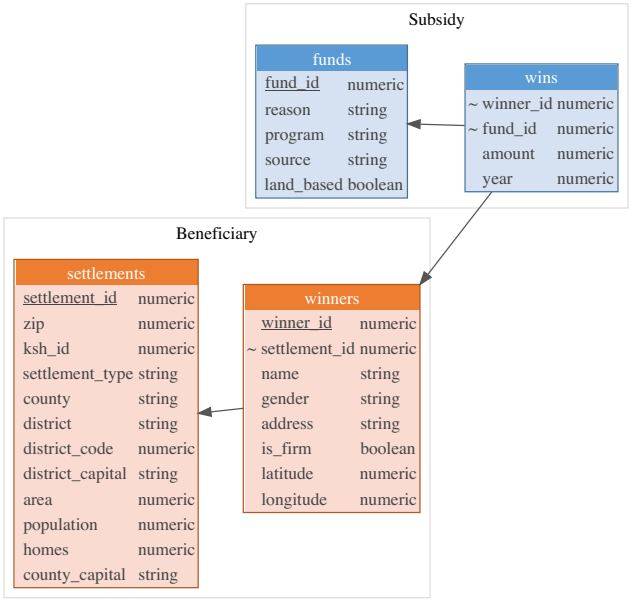


Figure 5: Denormalized schema of agricultural subsidies

The normalized dataset is split across multiple files but takes around half of the disk space as a single flat file.

```
## 40K ../output/agrar_funds.csv
## 354K ../output/agrar_settlements.csv
## 31M ../output/agrar_winners.csv
## 102M ../output/agrar_wins.csv
```

3 Data Exploration

The data contains 4 967 044 data points, each representing a single payment. If we aggregate over addresses we get the total amount of subsidies that a household or institution received. After doing this the number of records goes down to 1 691 404, which means beneficiaries on average received 2.94 payments each year. Unfortunately, the data does not allow handling of cases when people move but even with this inaccuracy the overall picture is much more meaningful if we do the aggregation.

3.1 Descriptive Statistics

The below table shows some descriptive statistics about the whole dataset. The range and standard deviation of subsidies are very large. The mean is more than 7 times as large as the median, which indicates that the distribution is skewed. The data contains 57 420 records where the amount is smaller than 1000 HUF and 355 786 entries where it is smaller than 10 000. The fact that these exist is a good indication of the inefficiency of government redistribution in general and the EU in particular. I would argue that amounts smaller than 10-20 000 HUF should be eliminated completely because the administrative overhead costs much more than any value these could create.

```
##            [,1]
## mean    1269935
## sd      9103160
## min           1
## max  3604348494
## p50      169638
## p95     4592753
## n       4967044
```

If divide the data to 10 bins (decils) and visualize the average and overall amount of money that goes into these bins we can get a sense of how skewed the distribution is. I also created a shiny app that allowed the client to play around with this 3D-view of the distribution.

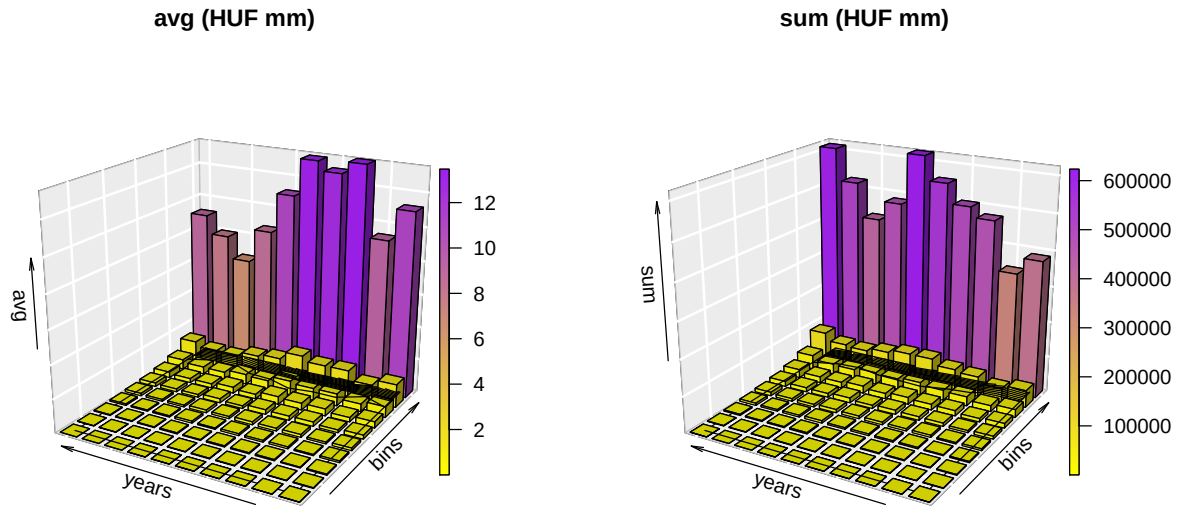


Figure 6: Average and sum subsidy by decils

The below table shows descriptive statistics for each year based on all payments. The distribution has similar characteristics for each year as for the overall dataset. It is interesting to see that the mean decreased in the past few years but the maximum amount almost doubled compared to 2010.

Table 1: Yearly statistics of all agricultural subsidies (K HUF)

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
amount										
N	344305	365242	351558	393167	417278	562007	555163	646040	657238	675046
mean	1424	1150	1696	1617	1728	1380	1158	869	1010	1188
median	189	135	216	201	222	157	206	141	149	181
sd	9682	7892	10819	10413	10239	10350	7534	6999	8422	9323
q1q3	68, 650	47, 463	76, 731	68, 722	76, 836	52, 536	72, 601	50, 424	46, 442	55, 598
min	0	0	0	0	0	0	0	0	0	0
max	1822327	1612995	1968611	1551840	2219240	3194617	3047052	2989005	3604348	3305794

The distribution is very skewed so visualizing the nominal amounts is not meaningful. The below violin plot shows the payments for each year using a logarithmic scale of base 10. So a value of 4 means a payout of 10000, while a value of 5 means 100000.

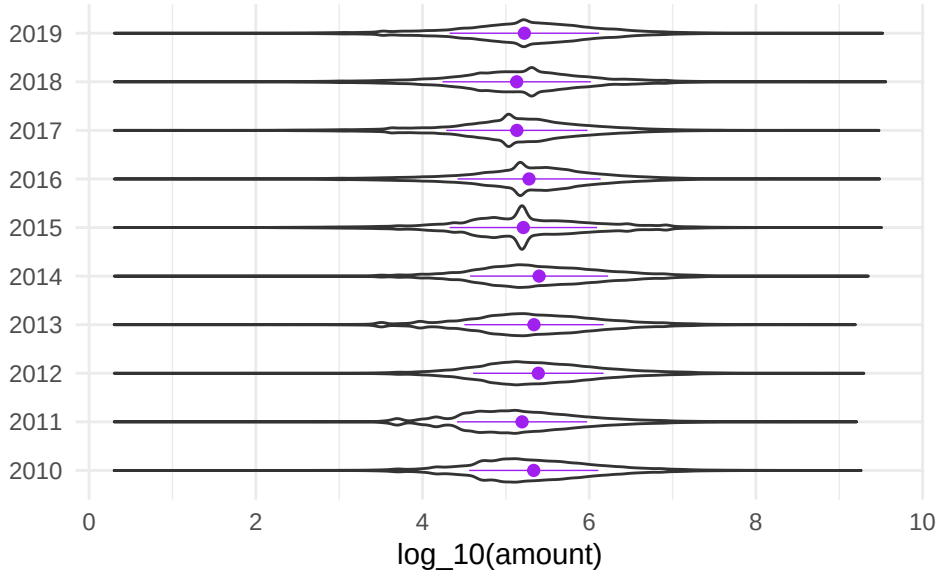


Figure 7: Yearly distribution of all payments

3.2 Subsidies by Address

When we look at the yearly distribution of subsidies summed by beneficiaries we see a steady increase in the mean and median values. If we compare the number of data points per year with the previous section we see 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 compared to earlier years.*

Table 2: Agricultural subsidies summed by beneficiaries (K HUF)

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
amount										
N	169757	167190	172538	175609	178509	178210	161400	163581	162472	162138
mean	2889	2511	3455	3619	4039	4352	3982	3431	4084	4945
median	306	201	349	375	427	480	588	412	440	675
sd	23608	21774	27435	28334	29224	31005	23838	22339	26132	29997
q1q3	114, 1062	74, 755	130, 1231	140, 1371	160, 1576	220, 1815	211, 1905	143, 1471	204, 1599	225, 2377
min	0	0	0	0	0	0	0	0	0	0
max	2981387	2515812	2211397	3590112	4126149	3725731	3047052	2989005	3604348	4458506

We can already see an increase in the mean if we look at the distribution on a violin plot.

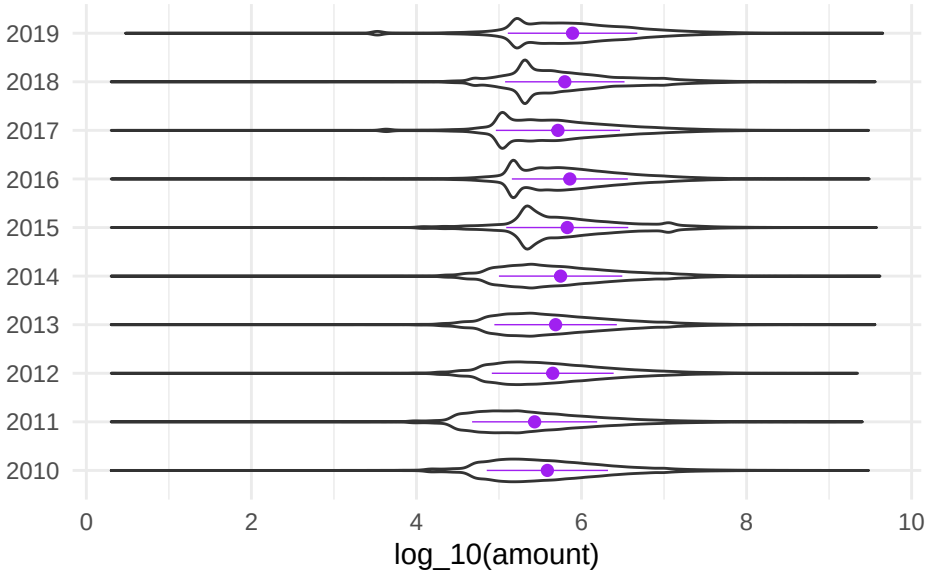
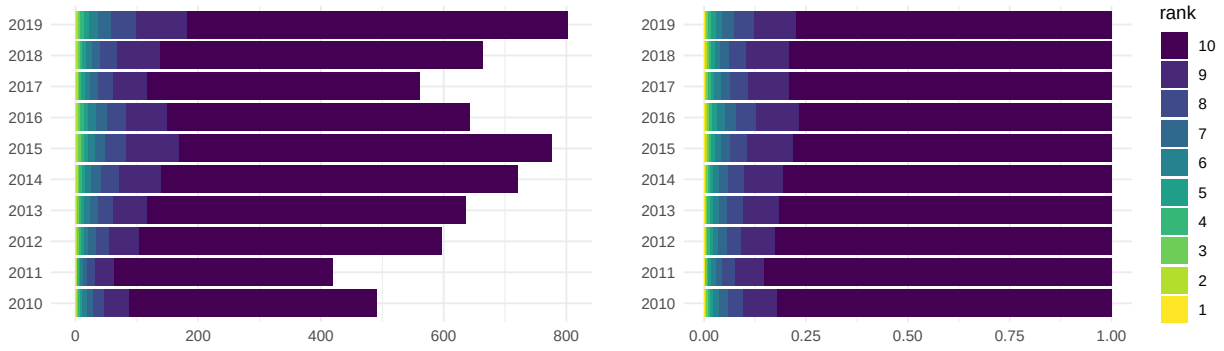


Figure 8: Yearly distribution of payments summed by address

The distribution visually is very similar to what the raw payments looked like. The top 10% accounts for more than 75% of the overall payments.



(a) sum amount (bn HUF)

(b) Percent of total amount

Figure 9: Distribution of individual beneficiaries by decils

On the log scale there is a steady increase between the bins except for the top 10% where the jump is around double of the others.

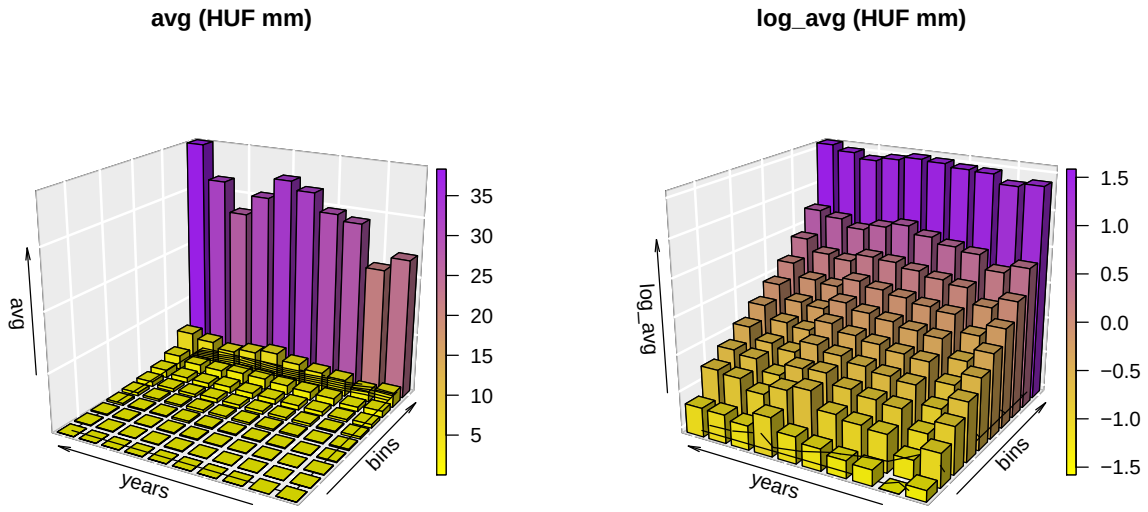


Figure 10: Average subsidy amount by decils

3.2.1 Individuals

The distribution is similar if we look at the subsidies won by individuals sharing the same address. In 2010 more than 163 000 households received agricultural subsidies while in 2019 it was slightly more than 154 000. Both the mean and median more than doubled during this time interval but the standard deviation also increased 2-fold. *It is notable that there is a household that received more than 525 million HUF in agricultural subsidies in 2019.*

Table 3: Subsidies won by individual beneficiaries (K HUF)

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
amount										
Count	162555	159412	163179	165595	167084	167033	154994	156974	155200	154367
Mean	1350	1063	1534	1708	1959	2190	2297	1940	2309	2727
Median	282	184	317	339	376	426	554	386	409	618
StDev	4394	3995	5074	5667	6653	6908	6204	5900	7363	8396
Q1, Q3	109, 913	71, 621	124, 1026	132, 1126	150, 1239	214, 1377	202, 1696	137, 1280	204, 1346	214, 2002
Min	0	0	0	0	0	0	0	0	0	0
Max	400630	271508	332788	470434	548516	556192	434562	356163	527485	525876

Both the average amount won by beneficiary and the total amount won surpasses the bottom 90% multiple times.

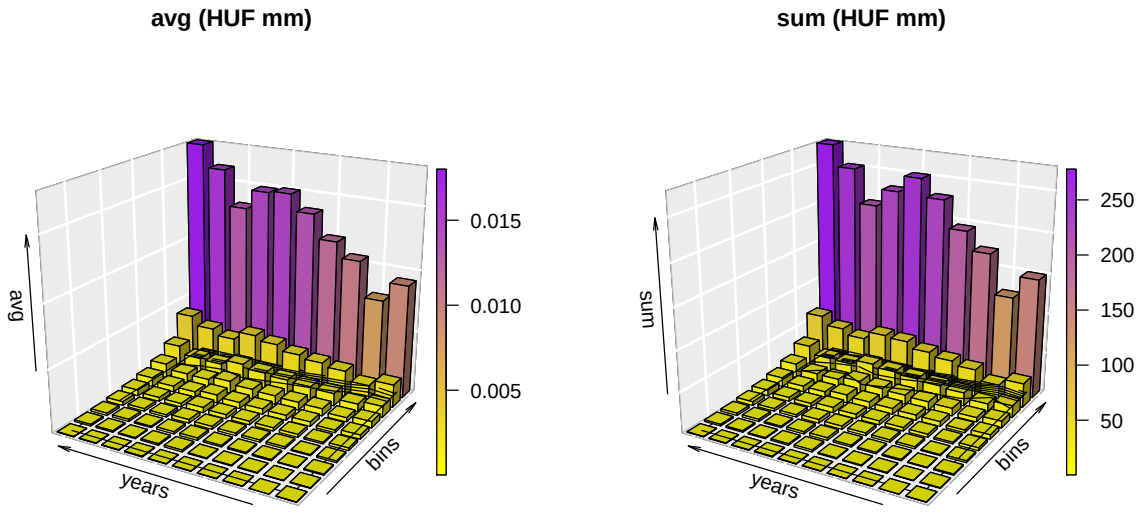
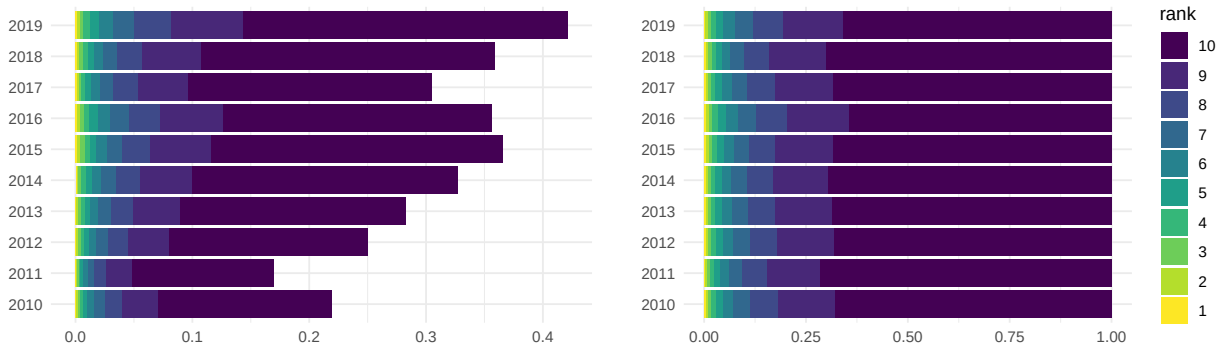


Figure 11: Average and sum of subsidies won by individuals

We see the same pattern that more money is distributed but to fewer participants and the top 10% takes the majority of the subsidies.



(a) sum amount (bn HUF)

(b) Percent of total amount

Figure 12: Distribution of individual beneficiaries by decils

3.2.2 Institutions

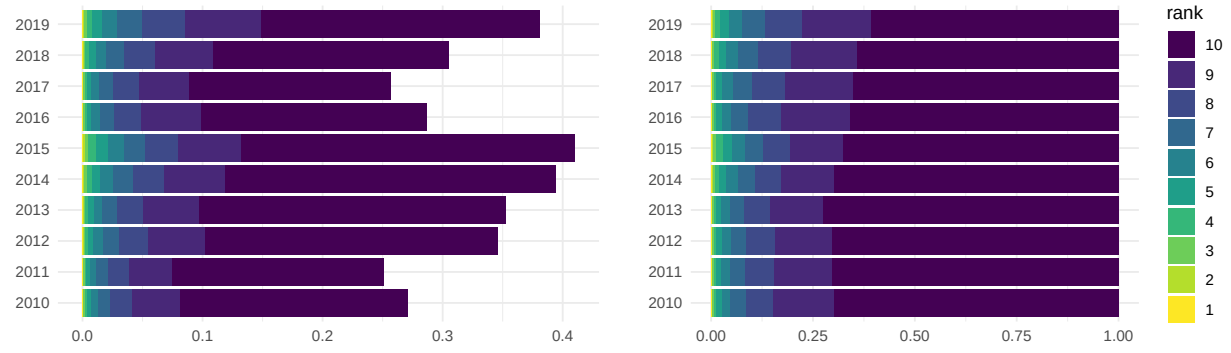
Table 4: Subsidies won by institutional beneficiaries (K HUF)

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
amount										
Count	9333	9802	11554	12372	13944	13679	8367	8707	9415	10000

	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019
Mean	29026	25551	29921	28512	28243	29956	34268	29482	32420	38073
Median	4846	4000	4978	4447	6089	8095	5977	5631	7500	9897
StDev	94804	84926	100170	101088	98198	105419	95535	88982	99452	110043
Q1, Q3	965, 19150	851, 17738	1029, 20956	1018, 17537	1539, 18521	2314, 19612	1081, 26507	961, 23829	1508, 26190	2207, 34772
Min	1	0	0	0	0	0	0	0	0	0
Max	2981387	2515812	2211397	3590112	4126149	3724586	3047052	2989005	3604348	4458506

When looking at the aggregate wins by institutions sharing the same address we see that the number of entities (rather addresses) who received agricultural subsidies was slightly higher in 2019 than in 2010. It is interesting that there was a steady increase in the number of recipients until 2015 when it reached almost 150% of the 2010 count. In 2016 there was a sharp drop and the number of institutional entities has been increasing since then. In the past 10 years the mean increased by 10 million to just over 38 million and the median more than doubled. *The largest sum won by a single firm was almost 4.5 billion HUF in 2019.*

If we divide the data to 10 bins of equal size the proportion of the top decil is striking in this dataset too. The proportion of the amount of money won by the top 10 percent of institutional entities has been slightly decreasing in the past 10 years but it is still above 60%.



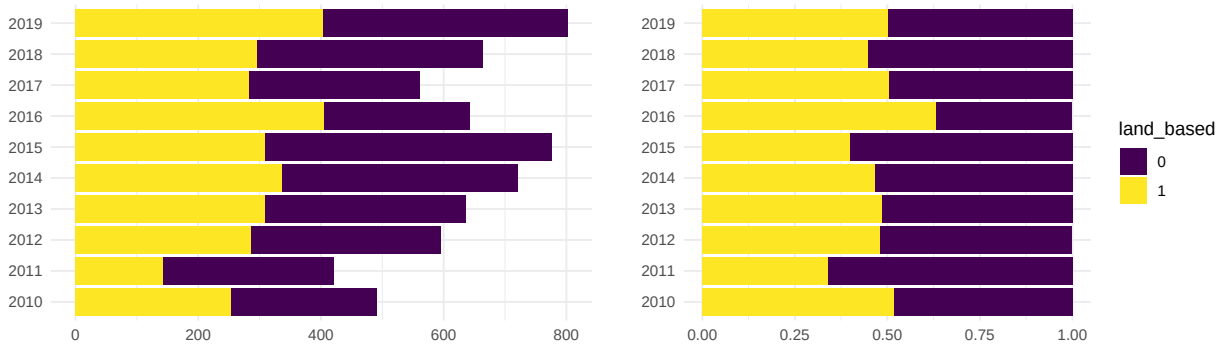
(a) sum amount (bn HUF)

(b) Percent of total amount

Figure 13: Distribution of institutional beneficiaries by decils

3.3 Land-based Distribution

About 50% of the subsidies are given to farmers just for their land ownership.

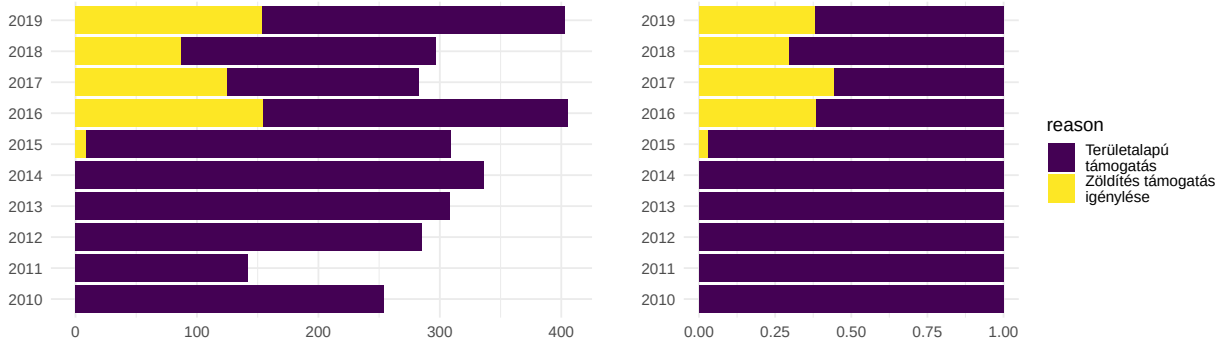


(a) sum amount (bn HUF)

(b) Percent of total amount

Figure 14: Proportion of land-based subsidies

Land-based subsidies were mostly given under a single category until 2016 when EU regulations changed and funds became available under another title.



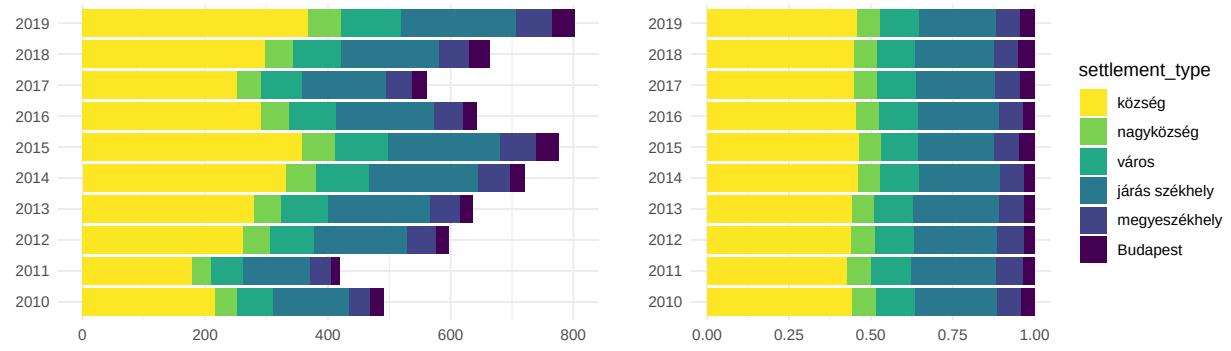
(a) sum amount (bn HUF)

(b) Percent of total amount

Figure 15: Composition of land-based subsidies

3.4 Settlement Type Analysis

The amount of money won by beneficiaries who reside in various settlement types. *It is important to note that the address is the official address of the firm or individual who received the money, not the address of the land, for which the money was given.* Most of the actual land belongs to villages and small towns but the proporion of the subsidies received by entities who are based in the country capital or county capitals is quite significant.



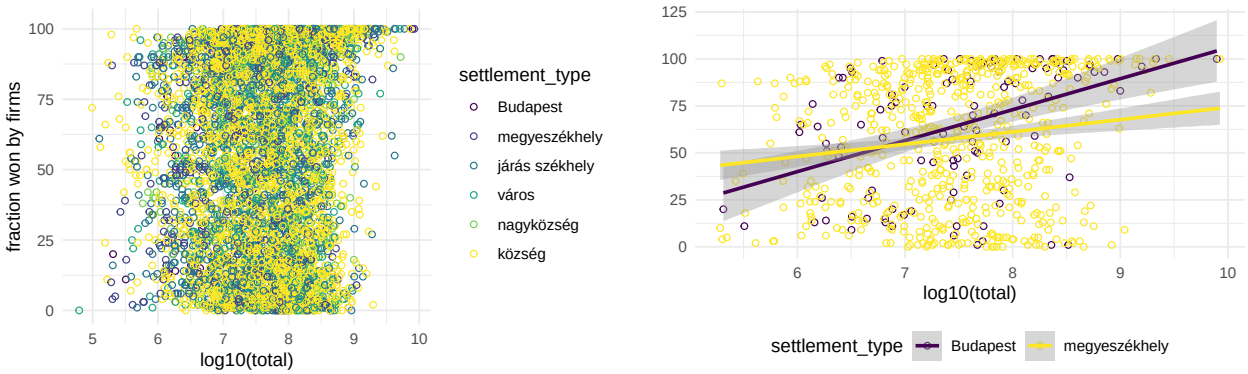
(a) Sum amounts in billion HUF

(b) Proportion of the groups

Figure 16: Subsidies based on the settlement type

3.5 Address Sharing

The total amount of agricultural subsidies distributed since 2010 is 6 307 822 959 247 HUF. From this amount 895 058 327 115 went to beneficiaries where individuals and institutions share the address. This is more than 14% of the total amount so it's worth checking if there is some pattern. The proportion of the amount of money received by firms versus individuals seems random for all settlement types except for larger cities, especially Budapest. In the capital the share of individuals is lower for higher total amount.



(a) All settlement types

(b) Budapest and county capitals

Figure 17: Proportion of subsidies received by firms when sharing address with individuals

4 Big Winners and Losers

Even though the agricultural subsidies data is publicly available the largest winners are not easy to find on the original website because searching for aggregates is not possible. The code I wrote resolves this problem. All data cleaning, transformations were done using public datasources so these findings could be extracted by anyone with enough skill and time.

4.1 Individuals

Here are the individuals who won at lease 1 billion HUF since 2010. Even though the names and addresses are available in the dataset I do not disclose those in this table. It is interesting to note that the first and third entries represent the same individual, who is the rock star of agricultural subsidies in Hungary.

Table 5: Individual winners with more than 1 bn HUF

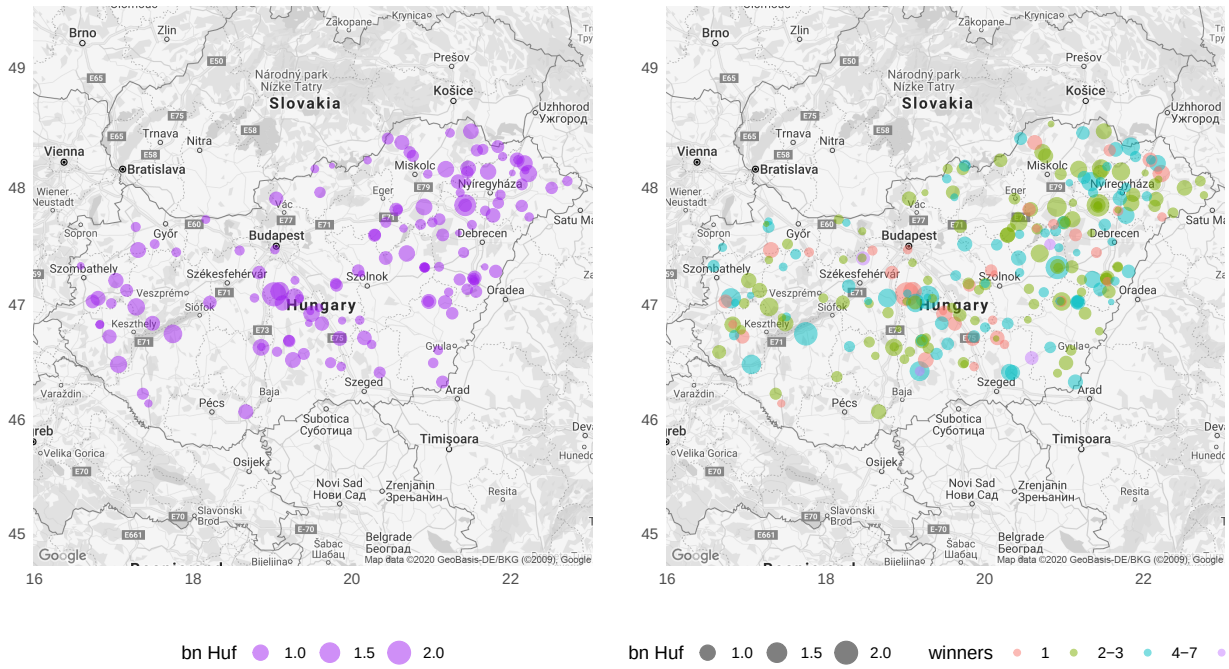
name	settlement	zip	total	num_wins
S. Dezső	Dömsöd	2344	2 039 943 435	104
N. András	Hajdúnánás	4081	1 657 920 157	55
S. Dezső	Apaj	2345	1 236 695 133	55
U. Zoltán	Látrány	8681	1 236 055 162	166
N. László	Kunpeszér	6096	1 185 344 679	107
T. Tamás Zsolt	Sümeg	8330	1 108 816 034	91
P. Károly	Ilk	4566	1 098 186 130	49
O. László	Nagyrécse	8756	1 074 546 360	69
T. Józsefné	Mezőcsát	3450	1 040 600 524	138
F. József	Paks	7030	1 036 382 303	140
C. György	Bekecs	3903	1 029 289 765	149
F. Szabolcs	Tomajmonostora	5324	1 007 759 942	119
S. Gábor	Kerekegyháza	6041	1 002 611 219	68

Below are the top items form the list created by aggregating wins by addresses. It is interesting to note that this list is also lead by an individual.

Table 6: Top 5 addresses by aggregate wins

settlement	zip	total	winners
Dömsöd	2344	2 039 943 435	1
Kunpeszér	6096	1 960 848 881	5
Látrány	8681	1 918 544 634	5
Karcag	5300	1 905 560 123	5
Hajdúnánás	4081	1 796 762 277	3

We can also put them on the map of Hungary.



(a) Single winners

(b) Individuals sharing the same address

Figure 18: Individuals who won more than 500 mm HUF

4.2 Institutions

Below are the largest institutional beneficiaries since 2010.

Table 7: Top 5 institutions by total subsidies

name	settlement	zip	total	num_wins
Atev Fehérjefeldolgozó Zrt	Budapest	1097	24 179 772 512	10
Bólyi Mezőgazdasági Termelő...	Bóly	7754	22 823 496 253	177
Nemzeti Élelmiszerlánc Bizt...	Budapest	1024	15 835 141 943	72
Agroprodukt Mezőgazdasági T...	Pápa	8500	14 140 332 882	174
Lajta Hanság Mezőgazdasági ...	Mosonmagyaróvár	9200	11 174 037 003	123

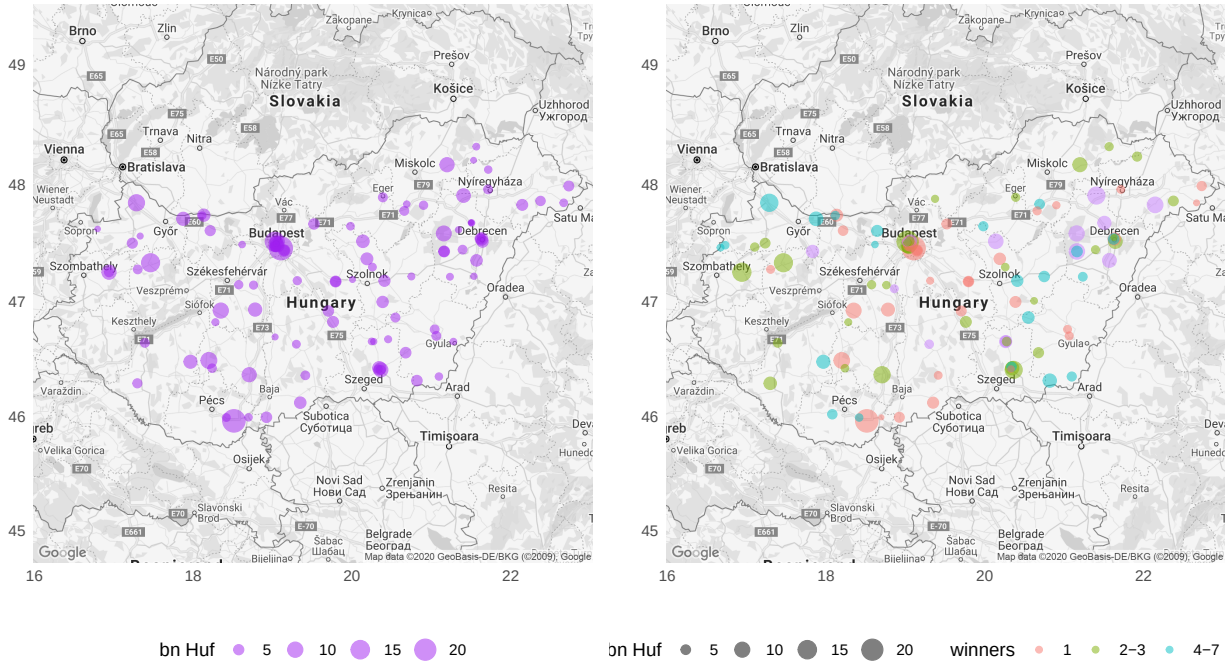
Some firms also share addresses with each other. Below are the addresses where the largest aggregate winners are located.

Table 8: Top 5 addresses by aggregate institutional wins

address	settlement	zip	total	winners
Illatos út 23	Budapest	1097	24 179 772 512	1
Ady Endre u. 21	Bóly	7754	22 823 496 253	1
Keleti Károly u. 24	Budapest	1024	20 079 984 014	3
Várkerület u. 26	Sárvár	9600	14 543 705 813	3
Szent István út 12	Pápa	8500	14 327 254 833	3

We can see where they are located on the map of Hungary. For individual winners Budapest was barely visible on the map but for institutions the capital is clearly a center. These institutions may perform valid

agricultural activity in the country but if their headquarters are in Budapest all subsidies they receive will appear in the capital in this dataset. More detailed analysis of the firm structures would help identify valid actors and freudulent players but that analysis is outside the scope of this project.



(a) Single winners

(b) Institutions sharing the same address

Figure 19: Largest 100 institutional winners

4.3 Address Sharing Individuals

Having a clean, tidy dataset allows running queries that would otherwise be impossible. One example query is below: what are the addresses that have the highest number of individuals who share the same address.

```
() xkey (select from
    (`cnt xdesc select name, cnt: count i,sum amount by settlement,address,zip from
      (select sum amount by name,settlement,address,zip from
        (select from .data.full where not is_firm)
      where address<>`)
  where cnt>5);
```

The below addresses would be worth checking for an investigative journalist.

Table 9: Top 5 addresses with the most distinct individual winners

settlement	address	zip	residents	amount
Kecskemét	Izsáki út 6	6000	49	272 639 837
Kecel	III. körzet tanya 77/1	6237	47	261 264 870
Székkutas	IV. körzet tanya 136	6821	41	767 229 913
Debrecen	Egyetem tér 1	4032	32	340 507 300
Baja	I. körzet tanya 7	6500	25	137 460 479
Nyíradony	Vörösmarty u. 26	4254	25	12 523 781

4.4 Settlements

We can also do aggregate statistics by county. The below maps show the yearly averages of land-based subsidies. If money were distributed evenly the circles and colors would be the same across the map. A larger circle means that the residents of the given area received more agricultural subsidies per unit area. Darker color represents a higher overall amount.

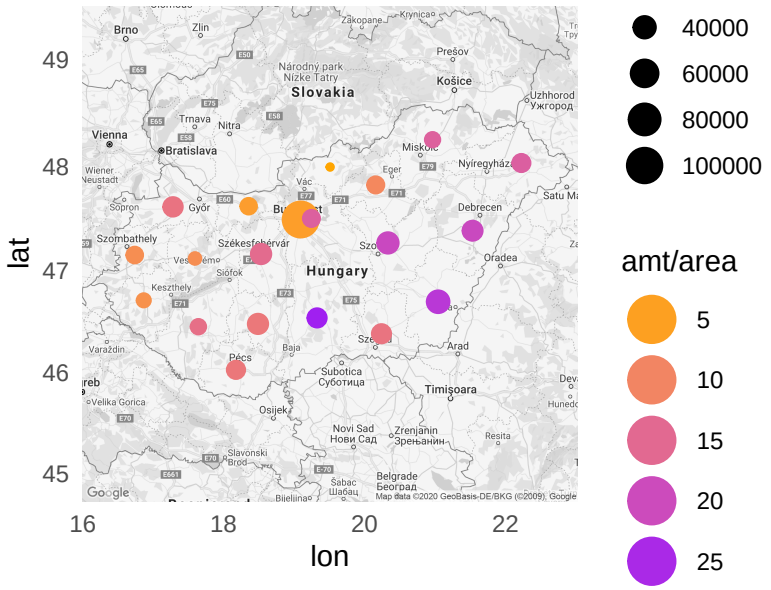


Figure 20: Land-based subsidies

4.5 Yearly Changes

One request from the client was to be able to view the largest winners and losers in a given area for each year by comparing yearly agricultural subsidies. The below function allows just that. You have to provide some constraint (eg.: filter for some area or subsidy type), the minimum yearly differenc and the number of largest records you want to see.

```
# .agrar.yearly_diffs[tbl:.data.full;constraint:(`land_based;(=;`zip;8086));limit:1000000;records:1]
.agrar.yearly_diffs:{[tbl;constraint;limit;records]
  yearly_amounts: ?[tbl;constraint;
    (`name`addr`year`zip`settlement)!(`name;(^;`address;`formatted_address);
    `year;`zip;`settlement);(enlist `amount)!enlist(sum;`amount)];
  winners: select distinct name,addr,zip,settlement from yearly_amounts;
  zeros: winners cross update amount:0 from select distinct year from yearly_amounts;
  diffs: update diff: deltas amount by addr,zip,settlement from
    select names: distinct name,sum amount by addr,zip,settlement,year from zeros, () xkey yearly_amoun
  diffs: delete from diffs where year = `$ string min "I"$ string exec distinct year from diffs;
  diffs: select from diffs where limit<abs diff;
  ret: () xkey (select from (`diff xdesc diffs) where ({[r;x]x in r#x}[records;];i) fby ([]zip;year)),
               select from (`diff xasc diffs)  where ({[r;x]x in r#x}[records;];i) fby ([]zip;year);
  xcols[`addr`zip`settlement`year`amount`diff;ret]
  };
```

5 Subsidy Dependence

Together with the client we talked to industry participants who highlighted some nuances that we wouldn't have otherwise been able to find out. The current system of subsidies is wrong for many reasons:

- it weakens market forces that would empower agricultural participants of the EU to compete globally
- Land-based sibsidies often land in the pockets of investors who do not have anything to do with agriculture, not the people who work the land
- Around 50% of the sibsidies are distributed following a selection processes where the subjective component is significant. This makes freud and corruption part of the system

Agricultural subsidies should at most be a supplement to the income of market participants, not the majority of it. For individuals it would be impossible for a non-state actor to analyze the financials because of the lack of financial data. For firms commercial datasources exist so the most important balance sheet components can be used to find what firms depend too much on subsidies and we can build models to categorize these. I got hold of some high-level financials of firms for 2018 and I could join these with the agricultural subsidy dataset. This allows the cration of statistical models to find non-trivial connections between the variables.

5.1 Model preparation

I defined two variables to measure how dependent a firm is on agricultural subsidies: a continuous and a categorical. The continuous variable is called `log_subsidy_per_sales`, which is the ratio of agricultural subsidies won by the firm compared to its total sales. I use a logarithmic transformation because the simple fraction is very skewed. The log values follow a normal distribution.

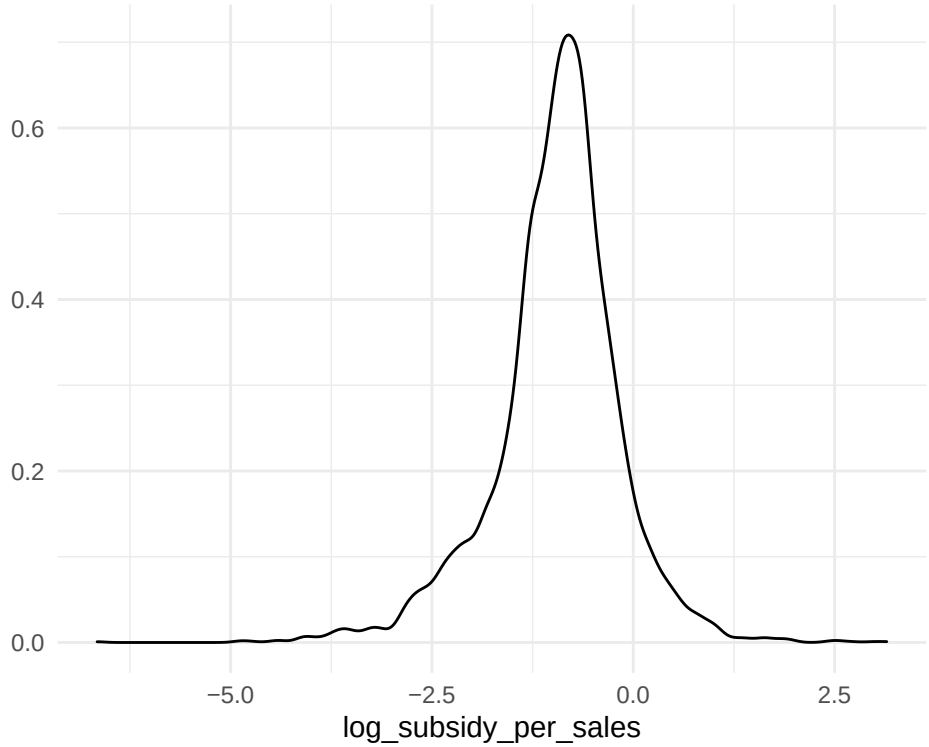


Figure 21: Density of subsidy_per_sales

There are dozens of descriptive variables that can be used for predicting the target variables but some of them cannot be used because they have a well-defined relationship with the target variable.

5.2 Linear Models

I created 4 models to estimate the dependence of a firm on subsidies: simple linear fit, Ridge, LASSO and Elastic Net. I use 10-fold cross validation in all of them. I use `caret` to estimate the models; this lets me compare them to each other using the same hyperparameters and folds. The RMSE of the linear model when looking at the holdout dataset is 0.6913906.

For the Ridge model the weights of descriptor variables tend to 0 as we increase lambda, the penalty term.

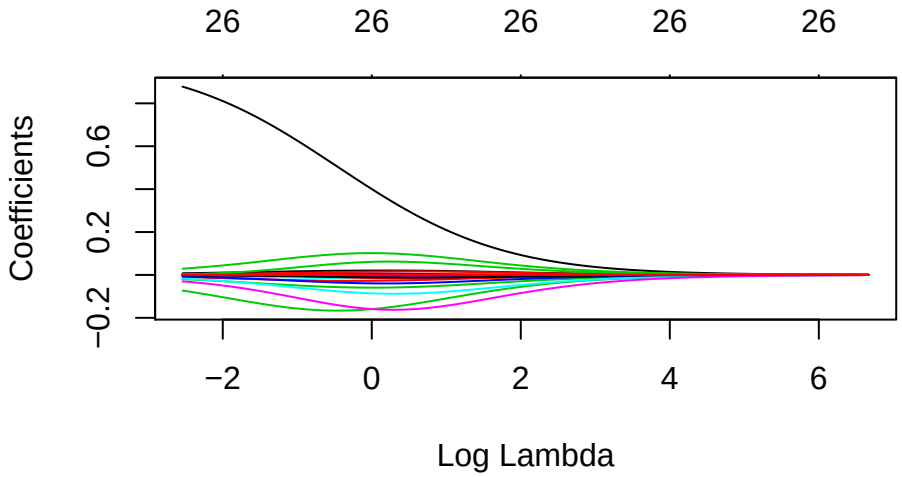


Figure 22: Variables fade to 0 as the penalty term increases

In penalized regression models Ridge shows a steady decrease in the RMSE as we increase the penalty term until we reach the optimum at 0.1.

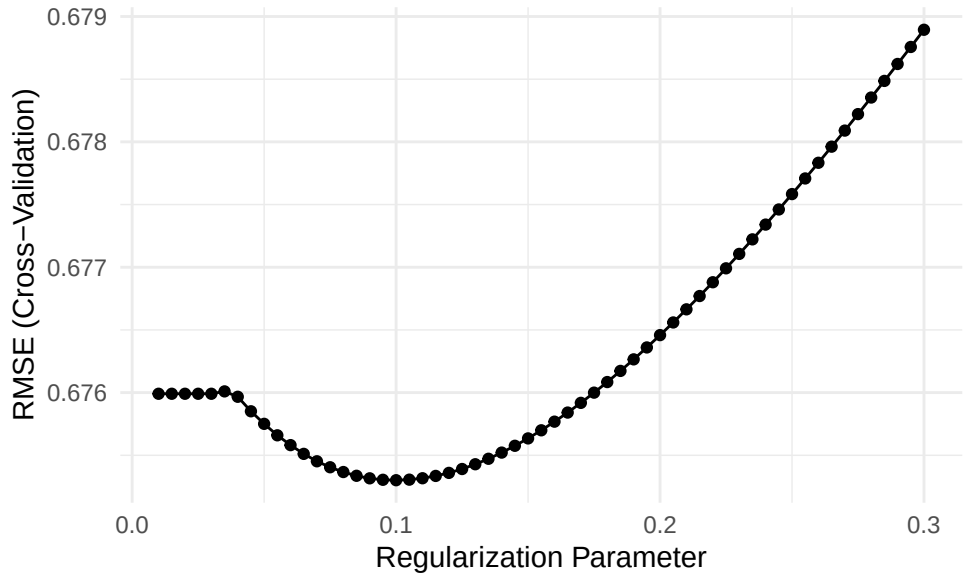


Figure 23: Ridge weight decay

The RMSE of the ridge model when looking at the holdout dataset is 0.6898465.

LASSO's RMSE looks different depending on the penalty term. The lowest RMSE is reached at 0.0316228.

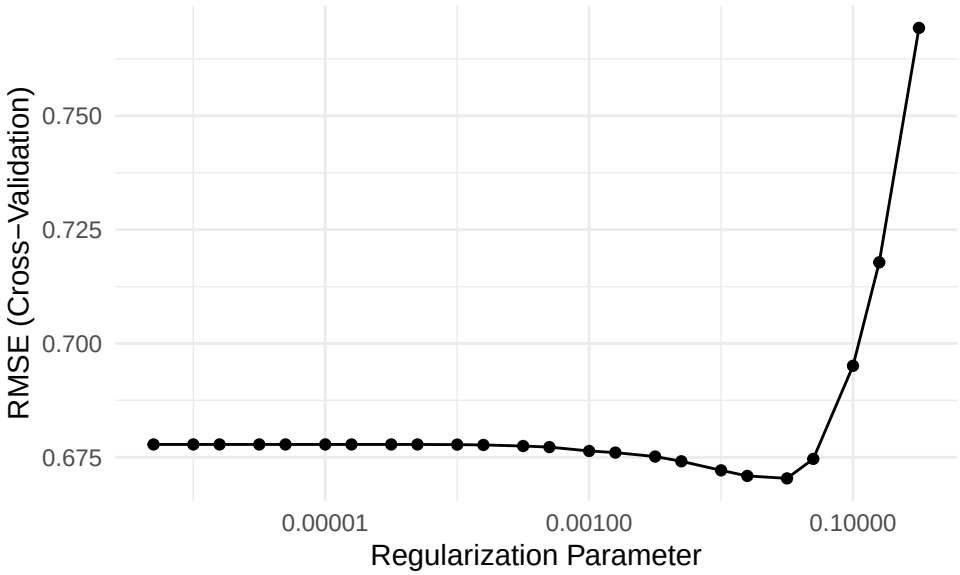


Figure 24: RMSE vs lambda for LASSO

The RMSE of the ridge model when looking at the holdout dataset is 0.6868989.

Elastic net combines Ridge and LASSO. We can see how the model performs for different alpha and lambda parameters. The best fit has an alpha of 0.8 and a lambda of 0.0316228.

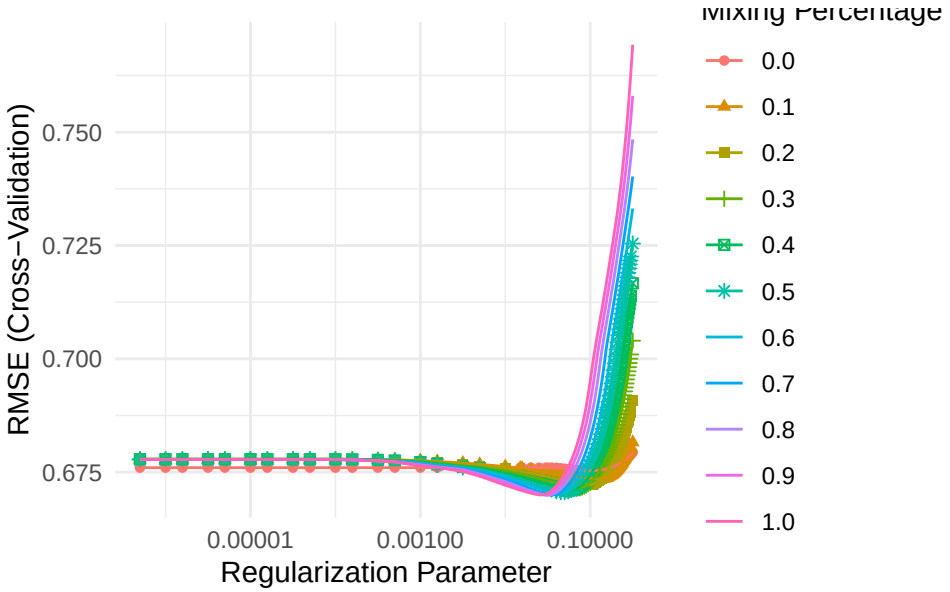


Figure 25: Elastic Net model performance for different lambda and alpha parameters

Simple linear regression has the highest error but the penalizing models have similar performance characteristics. All linear models with penalization perform slightly better than the linear model, elastic net being the best;

it has an RMSE on the holdout set of 0.6866017.

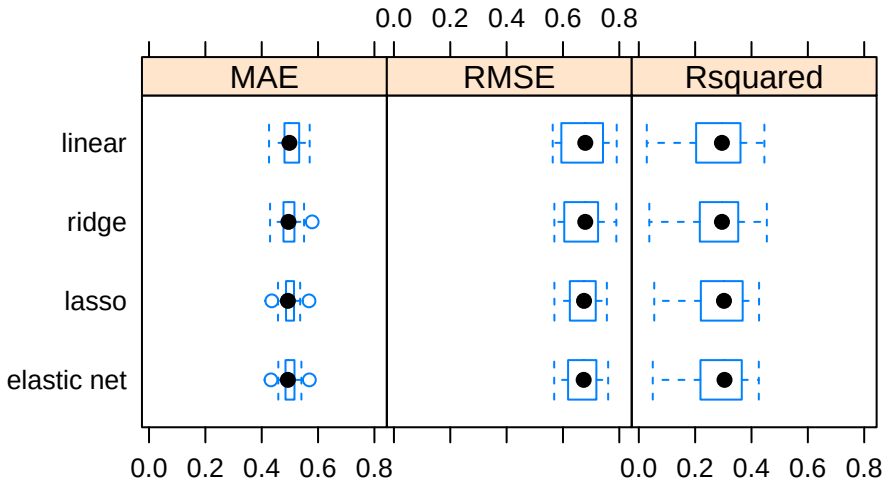


Figure 26: Linear regression model performance summary

We can use these penalized models to get those descriptive variables that can account for the largest chunk of variation in the target variable.

```
##                         variable    coefficient
## 1                    (Intercept) -0.94242638938
## 2             settlement_type.L  0.00705961515
## 3             settlement_type.Q  0.00000000000
## 4             settlement_type.C  0.00000000000
## 5             settlement_type^4  0.01157021297
## 6             settlement_type^5 -0.00007433321
## 7                       log_tax  0.00000000000
## 8             log_sales_per_emp -0.31610410997
## 9              pretax_per_sales  0.00000000000
## 10                amt_by_homes  0.08794968728
## 11                         emp -0.14357901181
## 12                  avg_salary  0.03224742122
## 13     a_eu_land_based_fraction  0.00000000000
## 14 a_eu_not_land_based_fraction  0.07142579406
## 15          a_national_fraction  0.00000000000
## 16   state_and_local_gov_owned1  0.00000000000
## 17                state_owned1  0.00000000000
## 18              foreign_owned1  0.00000000000
## 19             domestic_owned1  0.00000000000
## 20             local_gov_owned1  0.00000000000
## 21                  amt_by_area  0.00000000000
## 22                   amt_by_pop  0.00000000000
## 23                   population -0.03035984465
## 24                         area  0.00000000000
## 25                        homes  0.00000000000
```

The models do not have strong descriptive power and their external validity is also weak. The model does not

say anything about individuals but it would be quite interesting to see how dependent they are on subsidies. Even from the firm data a lot of data points were dropped either because they could not be matched with the financial results dataset or because of some missing values among their finances. I could use around 50% of the data points. Another factor that is not taken into account is that the analysis was done on 2018 data. Changes in political affiliation of an area and changes in regulations can change how firms treat agricultural subsidies.

5.3 Classification

We can also create a boolean variable from the "subsidy to sales" ratio by defining a threshold. It seems reasonable to say that 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 call this new variable `subsidy_dependent`.

```
##   no  yes
## 4101  755
```

For the classifier model I use 5 variations of logit trained on different sets of the available variables. I also trained a CART and a random forest model.

The best performing logit model is the one where all of the available variables are used. The AUC value on the train set is 0.8919909. On the holdout set it is 0.8549835.

The next classifier I built was CART, a simple tree-based approach. It is great because of its interpretability. At each split the algorithm maximizes the reduction in the error metric of the classification. The AUC value on the holdout set is 0.7671671. Each decision can be later examined and the most important variables collected.

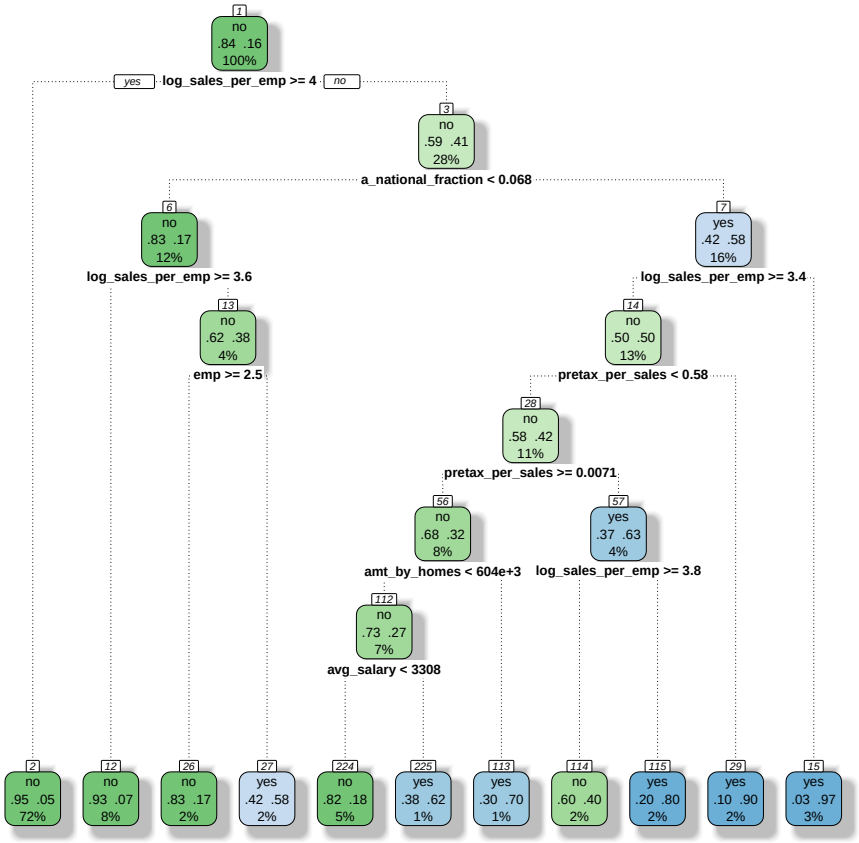


Figure 27: Decision tree for classifying subsidy-dependent firms

Random forest is a more robust, tree-based ensamble algorithm. It creates a large number of simpler trees using a number of of tricks, including bagging and using only a subset of the descriptive variables in each tree. Interpretability is more limited but the overall performance is typically better than that of CART. In our case the AUC is 0.8877251. We could inspect each tree but it would not give too much information because the strength of the model is in the numbers. However, we can aggregate the improvements in the classification for each variable in each tree to get the order of the most important variables. The top items in the list are usually the ones that appear in the CART model as well.

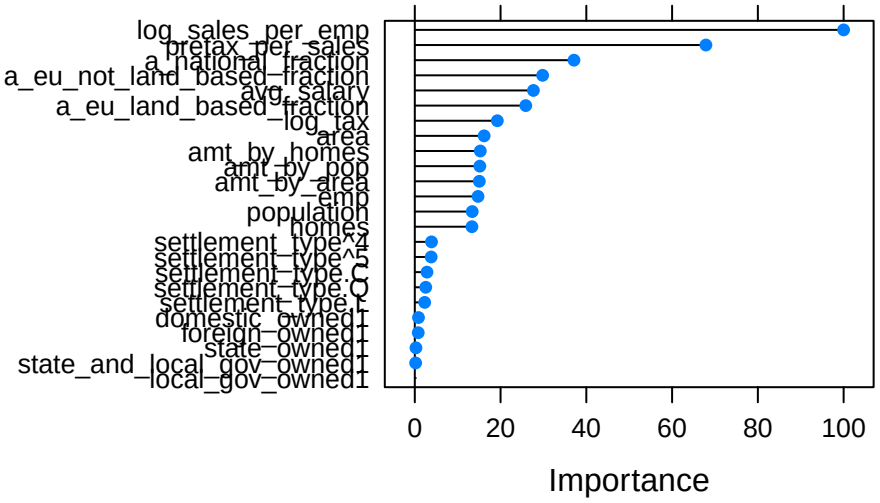


Figure 28: Variable Importance plot for Random Forest classifier

```
##
## Call:
## summary.resamples(object = .)
##
## Models: best_logit, cart, rf
## Number of resamples: 5
##
## ROC
##                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## best_logit 0.8254065 0.8612121 0.8670732 0.8730187 0.9018489 0.9095528    0
## cart       0.7783537 0.8273737 0.8281504 0.8345254 0.8558222 0.8829268    0
## rf         0.8329268 0.8723232 0.8943745 0.8916322 0.9154472 0.9430894    0
##
## Sens
##                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## best_logit 0.9207317 0.9454545 0.9634146 0.9586031 0.9756098 0.9878049    0
## cart       0.8841463 0.9212121 0.9329268 0.9354619 0.9573171 0.9817073    0
## rf         0.9634146 0.9634146 0.9878049 0.9780636 0.9878049 0.9878788    0
##
## Spec
##                 Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## best_logit 0.3000000 0.3225806 0.4333333 0.4045161 0.4666667 0.5000000    0
## cart       0.3666667 0.4516129 0.5000000 0.4969892 0.5333333 0.6333333    0
## rf         0.2666667 0.3548387 0.4333333 0.3909677 0.4333333 0.4666667    0
```

We can also visualize the ROC curve for each classifier. Random forest has the best performance while CART is below both RF and the best logit model.

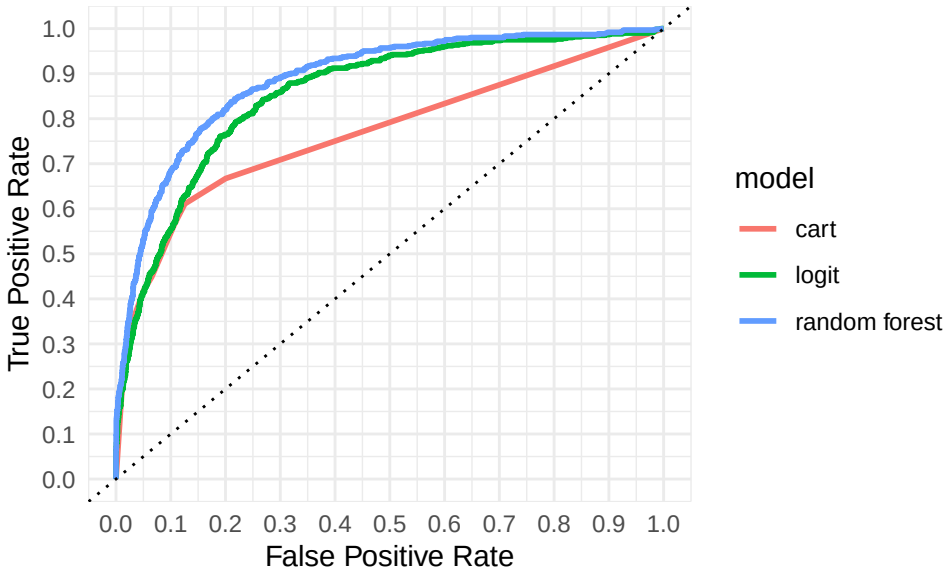


Figure 29: ROC curve of the classifiers

6 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 took first leaps to solve this by taking the raw data for the past 10 years for Hungary from the website of the Hungarian Treasury and applying a lot of data engineering techniques. I only used free tools and made the code available on github. 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.