Master's in Data Science Public Project Presentations.

On November 1st a multi-disciplinary team of masters' students presented at the PLNT Centre for Innovation and Entrepreneurship to representatives of the 2nd chamber of the Dutch parliament. This was part of an initiative organized by Professor Mirjam van Reisen and Mustafa Kedilioglu to showcase the potential of data science and data visualization for the various government departments. The students were from the Introduction to Data Science programme, a master's level course uniting students from both Computer Science and Statistics.

Predicting Conflict in Uganda using Geo-Statistics

Data science masters' students at Leiden University have used statistical techniques including Bayesian-kriging mapping and random forest classification algorithms to analyze conflict data from Uganda and generate predictive models.

This led to a diverse team with members from 4 countries, and backgrounds including Law, Psychology, and Economics, as well as Computer Science and Statistics. The team were provided a sizable Armed Conflict Location & Event Data Project (ACLED) data set compiled by the Bureau of Conflict and Stabilization Operations (CSO) containing Ugandan conflict data collected from independent media reports between 1997 and 2019. With this data the team was given the freedom to analyze the situation and gain deeper understanding of the conflict.

"I liked how this project gave me the opportunity to work with a great team of people on a topical, real-world issue"
Laura Jansén-Storbacka
Graduate Student in Statistical Science

Background to the Ugandan Conflict
Conflict in Uganda is a multi-faceted issue and summarizing the situation as a single conflict does not do justice to its true complexity. Uganda has many stress factors that can potentially lead to conflict. These factors include a large, young and fast-growing population combined with high rates of poverty and poor health care provision. There are a large number of diverse ethnic groups, and a series of unelected dictators has meant decades of political violence and government oppression.

Figure 1. United Nations Peacekeeping Forces Standing Guard (UN, Staton Winter)

Events classified as conflicts include riots and protests, organized violence against civilians, armed battles between government forces and militias, and strategic deployments involving peacekeeping forces. In North-Eastern Uganda, the notorious Lord's Resistance Army (LRA) has conducted an ongoing guerilla campaign since 1987, killing thousands of civilians and abducting children to serve as child soldiers. While LRA activity has decreased since the UN Peacekeeping intervention in 2007, militant activities still continue at lower levels. Additionally, in the South-East conflict in neighboring countries has led to a burgeoning refugee crisis near the border, with 1 refugee entering the country every 5 minutes.

From Data to Understanding

The starting point was the data. The data set contained information about location, date, conflict type and number of fatalities. However just looking at a giant spreadsheet isn't a sufficient way to understand a complex situation. The team began by examining and cleaning the data set, e.g. by removing duplicates where 2 media outlets reported on the same conflict. They then researched the history and background of the conflict in Uganda, consulting with officials at the Dutch foreign ministry as well as with Mariam Basajja a PhD student from Uganda. Next they created a timeline of key political events to bring the data into perspective. It was then possible to visually match events to the periods of greatest conflict, and to sort and create a graph showing conflict types over time.

"Our core drive was not just to use the data to make quantitative predictions, but to link the patterns in the data to a story involving real humans."

Ruduan Plug

Graduate Student in Statistical Science

The team also looked at the conflict data geographically. Levels of conflict were grouped according to their cartesian coordinates. It was then possible to produce an event count heat map grouping the geospatial data into bins indicating the areas of highest conflict.

Conflicts were also grouped using the main actor as an indicator of causality. This provided a clearer picture of how the different conflict types were concentrated geographically. Conflict appeared to be concentrated in three main areas. In the North there were many fatalities and high levels of military conflict involving militias and armed forces. Repeated patterns of violence were also evident near the south-western border. Additionally, there were riots and political unrest near the capital Kampala and in the broader southern region.

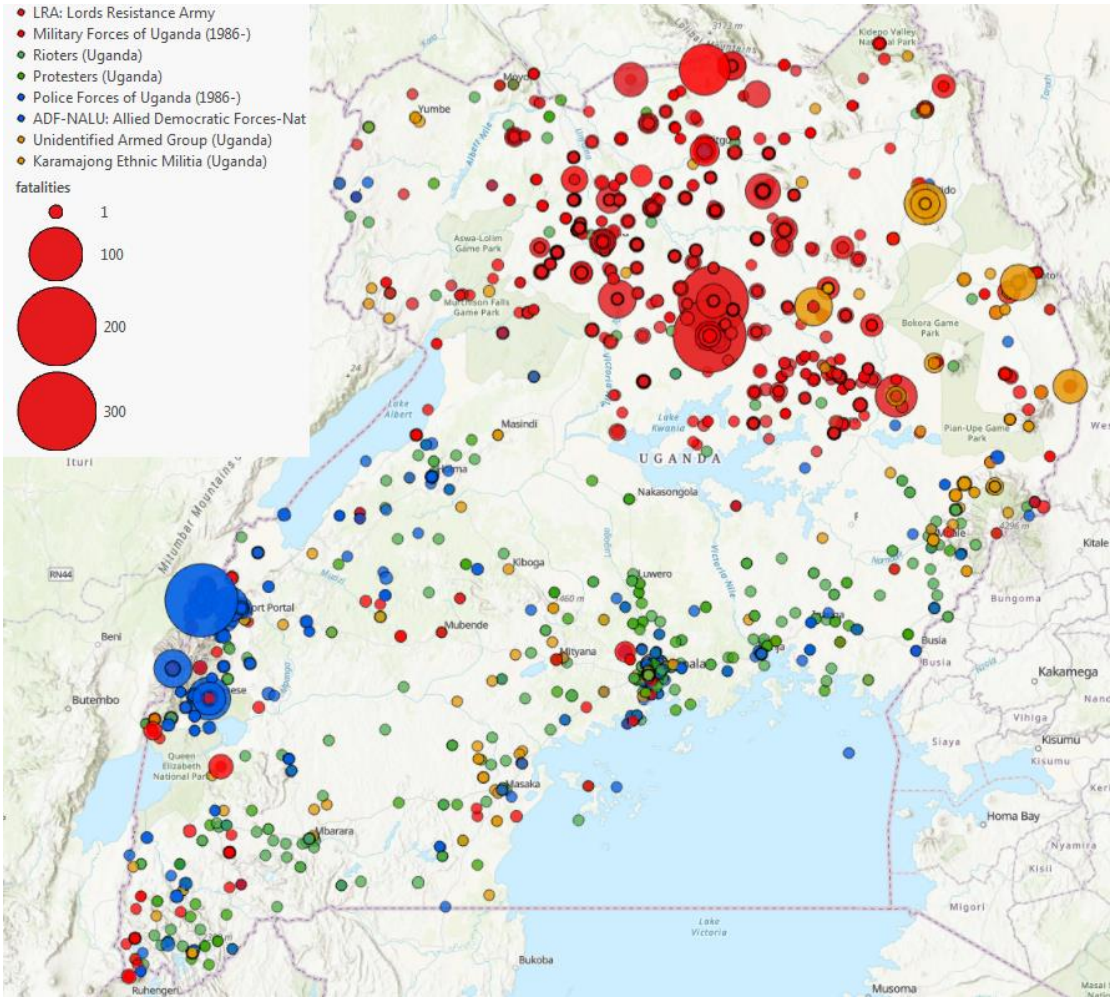


Figure 2. Geographic Distribution of Conflict Frequency Grouped by Initiating Actor
Red: Military Conflict, Orange: Terrorist Activity, Green: Protests and Riots, Blue: Police Action

A Probabilistic Conflict Model

Once the initial analyses had provided a visual understanding of the data, various predictive models could be constructed from the data. The team concentrated on two types of predictive model; Random Forest classification and Bayesian-Kriging techniques.

Random Forests algorithms were used to analyze our categorical data for predictive properties between each of the other variables. Random forest classifications use bagging –

bootstrapping using repeated random selections of the data – to create a forest of different decision trees. Many decision trees together allow a more accurate prediction than is possible with a single tree. This was used to evaluate the influence of various contributory factors to different types of conflict. The higher the influence of a factor on a particular type of conflict, the more useful it is in predicting it. Interesting findings included that violence due to protests and riots was mostly caused by the secondary, responding actor in the conflict – often government forces violently stifling peaceful protests, which would then turn into riots.

Bayesian-Kriging was used to create a dynamic geostatistical model showing how the probabilities of conflict changed throughout Uganda between 1997 and 2018, providing a measure over the stochastic nature of conflict. As more data points were collected over time, the accuracy in predicting conflict increased. These results were then combined into a predictive model where the statistical model was mapped onto the mesh of the map of Uganda. After being trained using the data from data from 1997-2018, the predictions of the model were tested using provisional data from 2019.

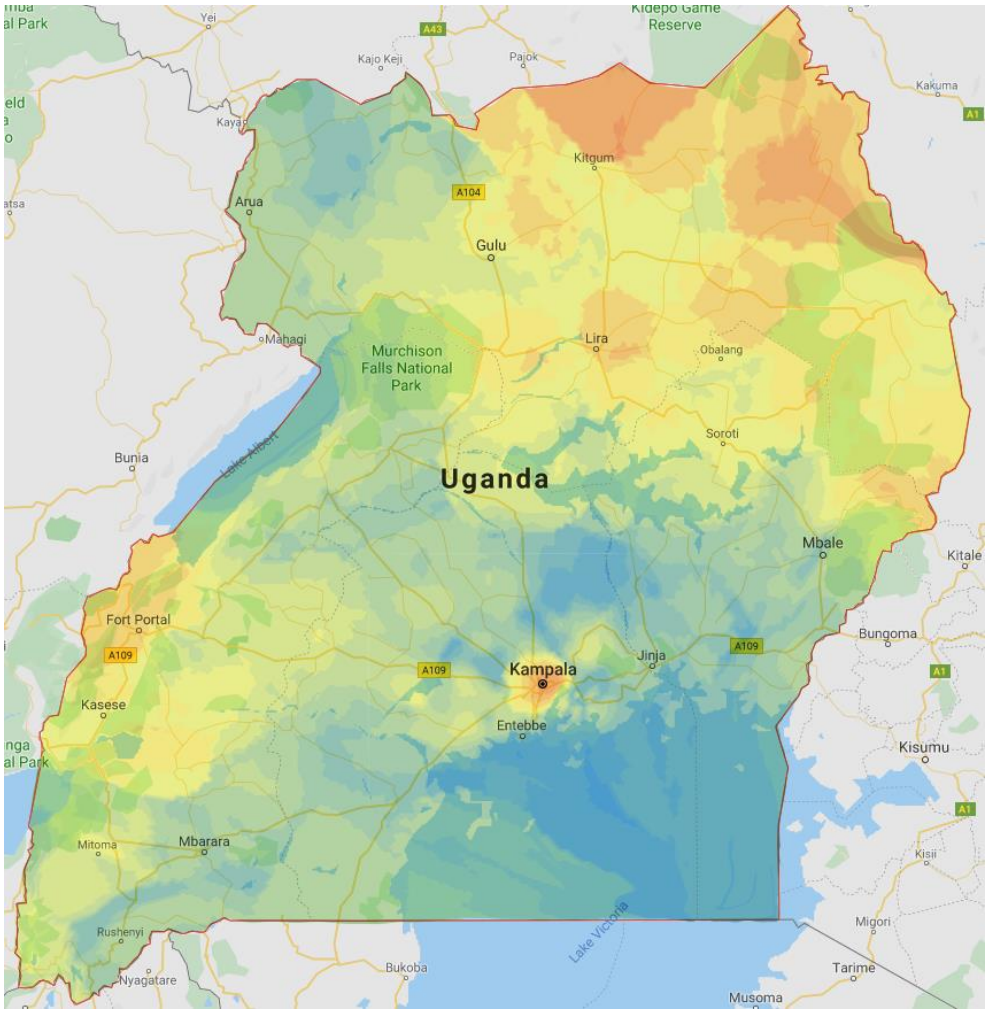


Figure 3. Final Geo-Statistical Model Showing Conflict Probability Densities in Uganda

The final model clearly displayed the three main conflict zones in Uganda, and visually communicated the probabilities of certain types of conflict in different areas of the country.

The team hope that in the future these models can help improve efficiency and specificity of peacekeeping efforts in conflict zones. *"In the future it could be possible to extend our geo-statistical model for different conflict zones, to directly aid peacekeeping forces and NGOs on the ground."* says Ruduan Plug. Concluding, Laura Jansén-Storbacka added: *"Bayesian-Kriging is a powerful geo-statistical method for generating probability maps; in one figure they tell a story that is worth a thousand words."*

WEEK100 TEMPORARY CASH FOR PERMANENT CHANGES

1. INTRODUCTION

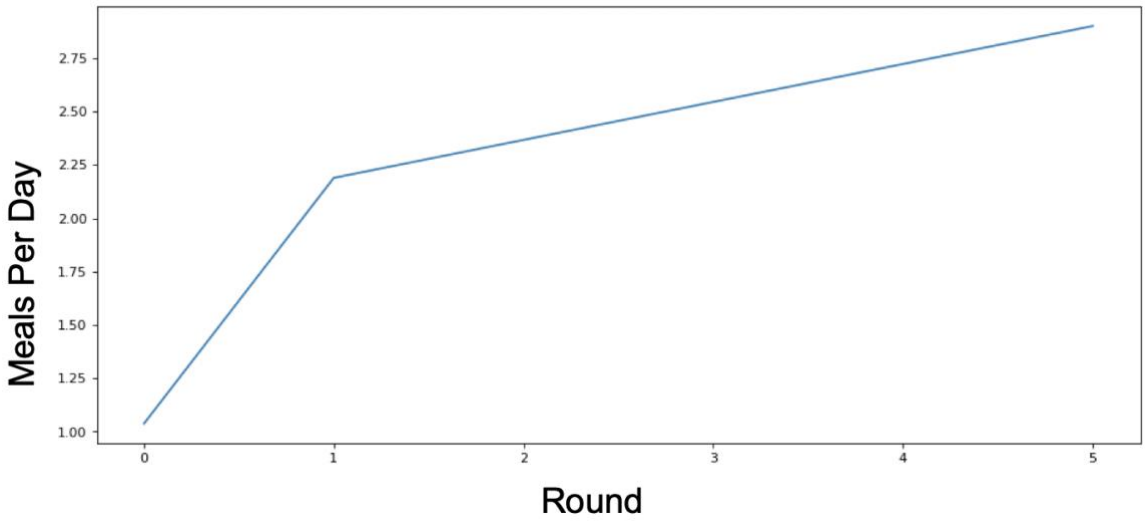
Week100 is a platform that sends money from contributors from Netherlands to women living poverty. It has been successful in helping women in Rwanda and is expanding its scope to other countries.
Large sums of money are put into poverty alleviation each year. It would be insightful looking into the data Week100 collected along the way to make the best use of the donations.
Week100 has provide data for several groups from 2016 in Rwanda. Rwanda is in central East Africa and has a significant problem of poverty. The data available includes the ages, education levels, household sizes, living conditions and etc. of beneficiaries in Rwanda. These data were collected for several rounds in a span of one year, during which Week100 would help and monitor if and how the beneficiaries improved their lives.

2. ANALYSIS

With the data we had, we would like to present the contribution of this project to improve the beneficiaries.
To assess this, we choose to analyse the results from the questionnaires on the happiness scale and the number of meals per day as a living standard index.
To measure improvement for the beneficiary, we performed a statistical test on whether factors such age, literacy or family size would influence the improvement of the happiness scale or living standard index.

3. VALUE OF WEEK100

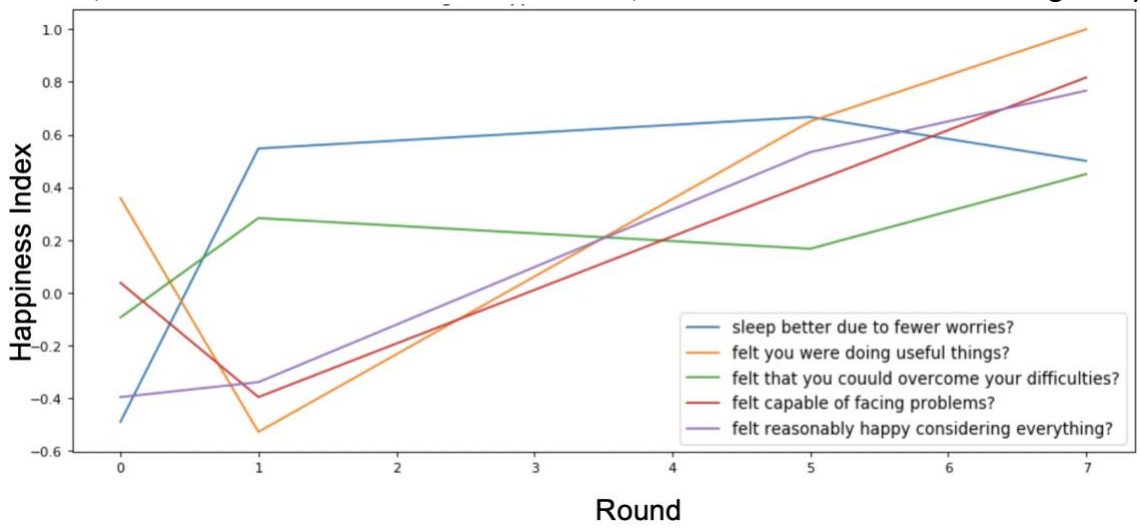
Meals per day is a simple number and can be used directly. We can see from the figure below that meals per day improved significantly from round 0 to round 5.



For the happiness index, we needed to preprocess the questionnaire answers from the beneficiaries. There are five questions regarding happiness:

- Sleep Better?
- Doing Useful Things?
- Can Overcome Your Difficulties?
- Capable of Facing Problems?
- Reasonably happy considering everything?

The answers can be 'Often', 'Seldom' or 'Never'. We assigned 1 to 'often' as the answers are positive, 0 to 'Seldom' as the answers are neutral, and -1 to 'Never' due to the negativity.



As we can see from the above figure, all the happiness index increased over the time. However, the index for 'Doing Useful Things?' and 'Capable of Facing Problems?' dropped for round 1. This is probably due to the beneficiaries taking new tasks and challenges after getting the money. It would be natural to be anxious and sometimes confused when taking on new challenges.

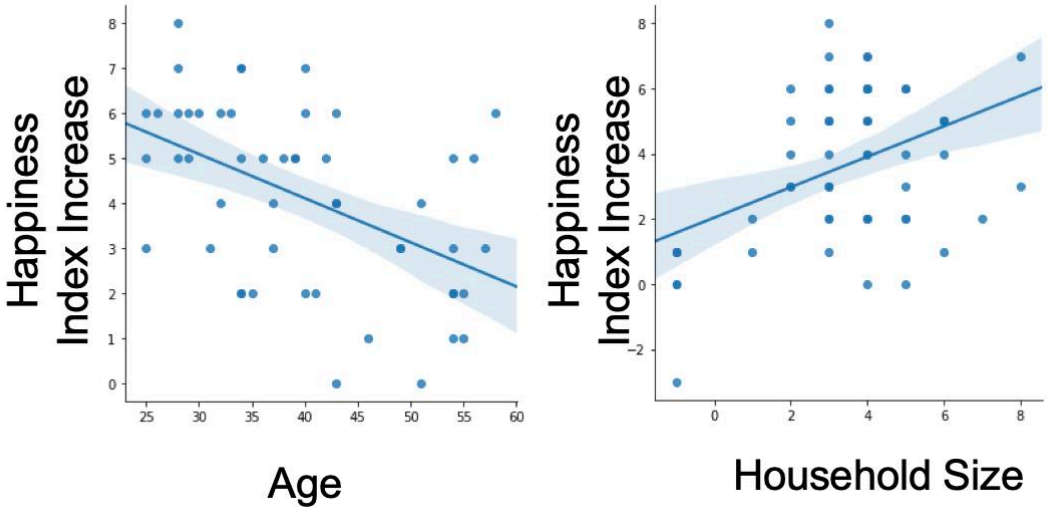
To test if the improvement is significant, we will also perform the before and after study. The p-values are close to 0 for both living standard index and happiness index. We conclude that the improvement for living conditions and overall happiness is statistically significant.

4. BENEFICIARY SELECTION

Besides seeing the benefits of Week100, we would also like to know what factors will influence the improvements. Since the funds are limited and we are not able to help all the women in target countries, knowing the critical factors can help select the beneficiaries for the future projects.

Based on the available data, we will test the factors of the age, the literacy, and the family size. The correlations between these factors and living standard as well as happiness index will be studied.

The results show that the influences of the ages and household sizes are statistically significant when it comes to happiness improvement. As shown in the figure below, the happiness index generally increases when the beneficiaries are younger or having bigger families.



5. DISCUSSION

We have found that both happiness and living standards improved for the beneficiary of Week100. Among these beneficiaries, younger women or women with more family members benefited more. We can definitely promote Week100 in more countries and test the platform first on younger women or women with bigger household sizes.

We have not found other correlations. But since our data is limited, there is the possibility that futures studies may show otherwise.

Also, the economy has grown rapidly in Rwanda these years. Since we do not have control groups, we cannot analyze how much the economy bloom contributed to the happiness and living standard increases.

Other problems we found is data inconsistency, and missing values. For examples, the ages changed significantly between rounds for two beneficiaries. It turns out some women did not know their exact age, and made an estimation. Also there are inconsistencies between school level and literacy. Staff from the field indicated that sometimes women are afraid to answer that they can read even if they can. This is due to the them being afraid that there might be a test on that.

There are also missing values for income sources, savings and etc. With more data, we can perform future analysis on these variables.

6. CONCLUSION

The 100 Weeks programme showed significant positive results on happiness and living standards. Happiness was measured through the following questions:

- Sleep Better?
- Doing Useful Things?
- Can Overcome Your Difficulties?
- Capable of Facing Problems?
- Reasonably happy considering everything?

The improvement in living standards were measured by the proxy of number of means a day that the person enjoyed.
These changes are showing that social-economic resilience has increased and that the objectives of 100 Weeks have been achieved.
It is recommended that the 100 Weeks programme is continued and that a number of improvements are made to advance the ability to measure progress of the recipients.

MAKE “HET RIJK” MORE ACCESSIBLE

During the last two months, our team of master data science students, analyzed the datasets of the Dutch government budgets and expenses. We met every two weeks, brainstormed at what angles we might apply statistics and in what ways we could interpret the outcomes, and assigned tasks for each of us to complete until the next meeting. Diversified backgrounds in this team, both educational and demographical, opened a lot of doors throughout the process. Professor Mirjam van Reisen and Mr. Mustafa Kedilioglu helped us formulate explanations that we, from a data perspective, would not have thought of.

The first problem we noticed was limited data access due to language. The “English” button on the upper-right of the parliament main page was practically useless; once one clicked it, the link to the datasets would be gone, making data barely findable for international researchers. Moreover, since data were presented purely in Dutch, Freek and Frederique, who are native Dutch speakers, had to spend hours explaining to the rest of us what every bit meant and which variables were useful.

We would like to suggest the implementation of a more efficient language tool, automatically translating non-sensitive information into English and simplifying data preparation. This move of improving the FAIR-ness should be crucial to projects, such as epidemics, that call for collaboration with neighboring countries, too.

Secondly, the labeling in the datasets was overcomplicated and yet not sufficiently informative. Most variables were named with four or more common words and abbreviated in a way that even confused native Dutch speakers. We had to guess the meaning backwards from values of that variable. In addition, there were multiple redundant columns, one of which was supposed to indicate the clustering of articles but basically just repeated article titles, increasing the volume of data without giving more information. Vague themes of each proposal gave rise to tremendous difficulty for categorization that will be discussed later.

For this labeling chaos, our advice would be eliminating redundant variables and using tags to extract features from each article. By doing so, politicians can input their expertise as highlighted keywords and help data scientists generate more interpretable results, which, in turn, will better support political decisions. Tags have been applied to many websites for decades and served as the starting point for plenty of massive data mining techniques, such as Locality-Sensitive Hashing.

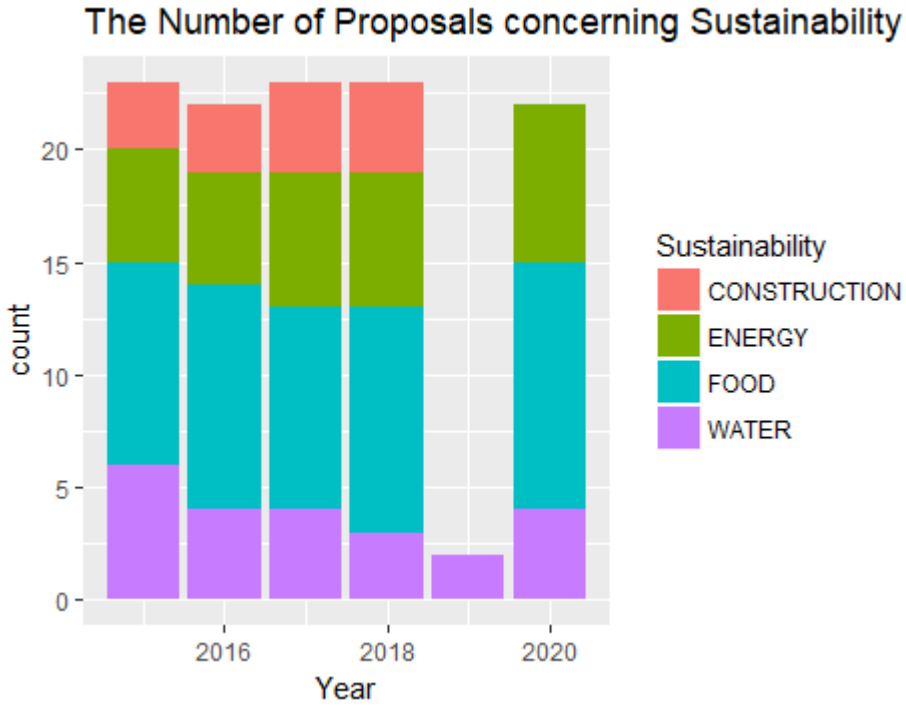
Finally, the data visualization displayed on the parliament website merely provided a high-level description of budgets and expenses, from which policy makers gain little insight. There are countless approaches to dive into massive data. For showcase purpose, we employed statistical methodology and explored “what people really care about”.

Our initial idea was that widely-concerned issues had probably been aware of and allocated resources to. Hence we started by picking out the five topics which received the largest amounts of budgets in 2019 and 2020. It turned out that in both years, the top five “hot” topics were finance, social affairs, public health, national debts, and municipal funds; and the

budgets did not differ much across years, meaning that the concerns at this level had equilibrated.

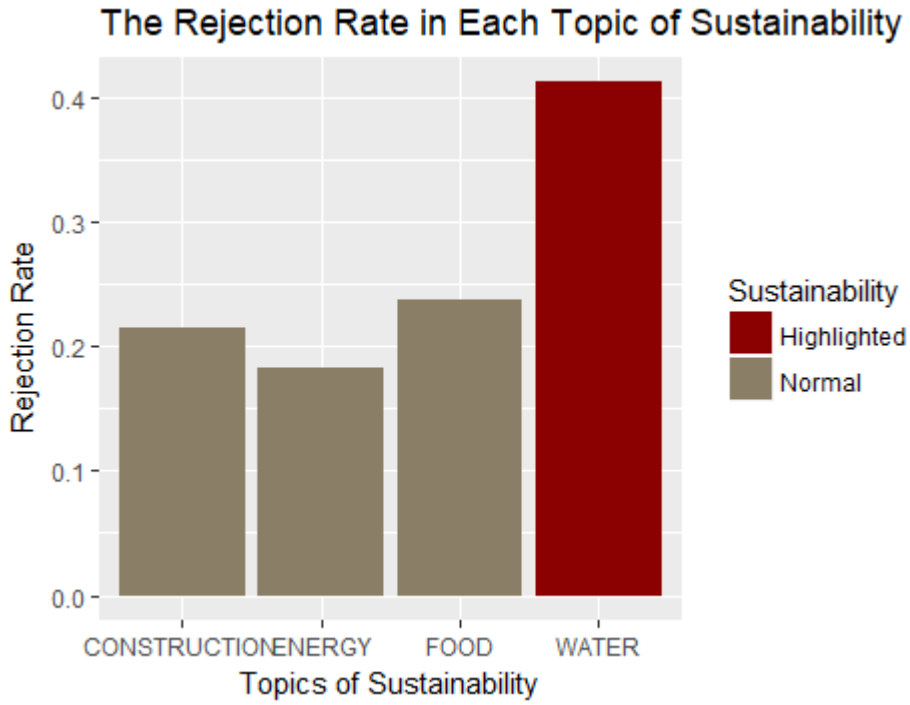
Among the top five, municipal funds attracted our attention in particular as it could be related to personal experience. It is very likely that nobody would notice a remarkable advance in finance; but people certainly would not miss a renovated soccer field in their neighborhood. It follows that tax payers would like to know whether their money is used wisely, or the municipalities just spend however much they want regardless of budgets. To answer that question, we conducted a compared T-test between the planned budgets and the actual expenditures of municipalities over the past four years. The good news is that no significant difference was found between budgets and expenditures, so the Dutch government did a great job at budget control.

Now that we had brought ourselves to the perspective of individuals, an intuitive assumption would be that if people really care about something, they should submit a proposal through their congressman and request funding. Therefore, instead of focusing on what has been actually spent, we looked into what people had asked money for. The globally popular topic, sustainability, was chosen, even though it was not among the top five, and divided into four sub-categories, namely, water, food, energy, and construction sustainability. A typical construction sustainability concern is the landfilling of construction waste. Then we visualized with a segmented bar plot the number of budget proposals in each sub-categories from 2015 to 2020, and this is where efficient keywords would greatly facilitate as mentioned above.



Clearly, data for 2019 and 2020 were problematic. There were no budget proposals concerning any kinds of sustainability other than water in 2019, and all of a sudden, those concerns came back in 2020 except for construction. If Mr. Kedilioglu was correct about the reason he gave to us, that is, parliament changed the names of topics frequently, a more consistent workflow would be highly recommended. On the other hand, data for 2015-2018 displayed homogeneity; that is to say, a Chi-Square test showed that distributions of budget

proposals with respect to four sub-topics only differed insignificantly across years, which allowed us to merge those data and dive deeper.



As the scope narrowed down, zeros in the budgets column became eye-catching. Zero budgets defined as being rejected, proposals concerning water sustainability had been rejected approximately twice as often as the others, as the bar plot of rejection rate in each sub-topic showed. It is possible that people worried about water overwhelmingly; many of the proposals were unnecessary or repetitive. Or there was simply not enough funding for water related funding proposals. In either case, stakeholders ought to be informed correctly. Furthermore, Mr. Kedilioglu proposed the third possibility that zeros could be caused by errors during data collection, which went through thousands of hands and the procedure was not designed for computer operations in the first place. In that case, singularities might not exist, but why errors occurred unusually often around water sustainability still deserves investigation. And we want to point out that automated data collection should be a less error-prone process.

Hopefully, the example above might give a concrete idea of what insights big data could yield and how the results could inspire decisions. It is our pleasure if the findings and advices would promote the accessibility of parliament datasets to open up transparency to the public to a much greater extent.

Analysis of Voting Records from the Dutch Parliament

Introduction to Data Science

For the first year MSc students in both Computer Science: Data Science and Statistical Sciences: Data Science, we participated in a mandatory course called Introduction to Data Science where the students had to develop a project. Some groups were assigned to the conflict data analysis on Uganda and others worked on projects for the Dutch Parliament.

Several topics for the project were proposed by the professors but our group decided to make a try for a different project and set out to analyze which categories of votes sparked the most controversy within political parties. Was it concerns regarding 'Society' or perhaps the 'Climate' that caused most Members of Parliament (MPs) to vote independently of their colleagues in their own parties.

Considering that the data regarding voting records that way we needed them were not publicly available in a processable file, we had to create our own dataset. Albeit real, it included a small amount of data that we had to manually extract from Parliament website and insert in a .csv file.

After an initial data analysis, we did not find that many votes that caused MPs to vote against their party line. Some votes where this happened included, for example, the proposal for citizens to be, by default, organ donors unless explicitly stated otherwise, but the votes where this happened in our small dataset were far too few compared to all the others to make a meaningful analysis.

As such, we decided to change our project and started to work with trying to find which votes caused the most controversy in the entire Parliament, that is, for each chamber[1], to find the votes that either barely passed and split the chamber in half.

[1] 1st Chamber as the Senate and 2nd Chamber as the House of Representatives.

To do this, we calculated the percentage of yes votes for every vote and then decided that the closest the percentage was to 50,0% then the most controversial the vote. Similarly, the closest the percentage was to 100,0%, the least controversial the vote.

The same could be done with no votes. However, considering that, for simplicity, we wanted a perfect 1 to 1 correspondence between our dataset from the Second Chamber and out dataset from the First Chamber, all votes had to pass the Second Chamber so, for this one, all votes in the dataset had a percentage of yes votes above the 50,0% threshold.

The top 3 of most controversial votes we found in the Second Chamber are given in the table bellow:

Vote description	% of yes votes
Automatic status as an organ donor	51,14%
Abolishment of referendums	52,00%
Vote on the gas extraction in Groningen	54,97%

The top 3 least controversial votes we found in the Second Chamber are given in the table bellow:

Vote description	% of yes votes
Removal of appointment of mayors by the King in the Constitution	98,01%
Introduction of compulsory military service for women	97,99%
Paris climate deal	95,36%

The top 3 of most controversial votes we found in the First Chamber are given in the table bellow:

Vote description	% of yes votes
Abolishment of criminal immunity for Government officials	49,33%
Automatic sharing of property and debt after marriage	50,67%
Extension of the naturalization period from 5 to 7 years	48,00%

The top 3 of least controversial votes we found in the First Chamber are given in the table bellow:

Vote description	% of yes votes
Introduction of compulsory military service for women	97,33%
Football vandals get area bans	92,00%
Treaty of Prüm	91,78%

Taking on these results, we counted the number of controversial votes that each category had and were able to create a sorted list of the most controversial categories for both the first and Second Chambers. The results follow bellow, being on the top the most controversial ones.

Second Chamber	First Chamber
Society	International Relations & Foreign Affairs
Security & Defense	Society
International Relations & Foreign Affairs	Security & Defense
Governing	Governing
Migration	Migration
Climate	Climate

It's particularly interesting, and reassuring, to see that for both chambers 'Climate' is the least controversial topic and that most MPs agree on how to proceed in this matter. As expected, Society, which englobes votes such as the burka ban, the law imposing openness about

artificial fertilization donor data, and LGBT issues, was found to be controversial in both chambers, toping the chart for the Second Chamber.

Additionally, we decided to try to work with similar votes and made use of the Jaccard similarity measure. Initially, we set the Jaccard similarity score to 0.50 in order for two votes to be considered similar, however, in a later stage and following the given advice on existing literature[2], we lowered it to 0.20, finding more relevant results.

First, we started by simply finding all the pairs of similar votes in the dataset, getting results such as the the association treaty between Ukraine and the EU (Vote 1) being found to be similar with the law that regulates how lawsuits from one EU member state should be enforced in another EU member state (Vote 2) and then the algorithm plots the results for both, as seen in figure 1.

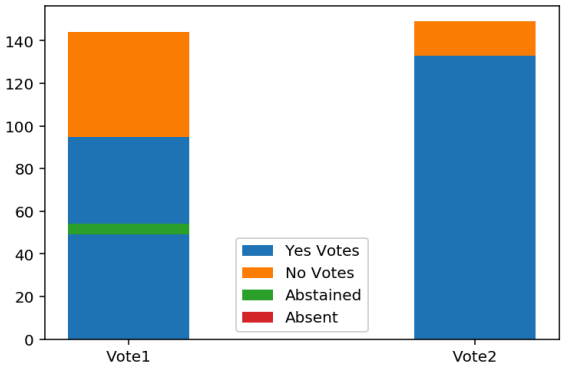


Figure 1: Results for two similar votes

Then, we can give the user the capability to, given keywords inserted by the user, provide information about similar votes that have happened in the past. For instance, if we are looking for previous votes that are similar to asylum and migration, we can type those keywords in the program, as per figure 2, and the algorithm will iterate through the database and use the Jaccard similarity to find similar votes.

Figure 2: Option to insert keywords and then find similar votes

For this particular example, the previous vote that was found to be of more relevance was the one where it was debated whether asylum seekers should not be given priority when looking for a home. It is also possible for the program to group all similar votes found and plot their results collectively on the same graph so the user can easily see how the Parliament usually votes on those matters.

---

[2] Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. Mining of Massive Datasets, 9781107015357, 1–315. Available at: http://www.mmds.org

Group Information

Team Leader and Contact Person

André Cardoso Silva Ferreira

a.f.cardoso.silva.ferreira@umail.leidenuniv.nl

Technical Development

Aaron Dunlea

Georgy Gomon

Canvas and Financial Analysis

Anish Kisoentewari

Sjoerd Hermes

FAIR and Presentations

Jonas Ammeling