ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Lecture with Computer Exercises:
Modeling and Simulating Social Systems with MATLAB

Project Report

Predicting Syrian Refugee Migration in Europe

Lukasz Pietrasik & Sylvia Schumacher & Wojciech Witon

Zurich
Dec 2016

Agreement for free-download

We hereby agree to make our source code for this project freely available for download from the web pages of the SOMS chair. Furthermore, we assure that all source code is written by ourselves and is not violating any copyright restrictions.

Lukasz Pietrasik Sylvia Schumacher Wojciech Witon

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of Originality

This sheet must be signed and enclosed with every piece of written work submitted at ETH.

I hereby declare that the written work I have submitted entitled

Refugee Modelling

is original work which I alone have authored and which is written in my own words.*

Author(s)

Last name	First name
Witon	Wojciech

Supervising lecturer

Last name	First name
Sanders	Lloyd

With the signature I declare that I have been informed regarding normal academic citation rules and that I have read and understood the information on 'Citation etiquette' (http://www.ethz.ch/students/exams/plagiarism_s_en.pdf). The citation conventions usual to the discipline in question here have been respected.

The above written work may be tested electronically for plagiarism.

Zürich, 12th Dec 2016
Place and date

Wojciech Witon
Signature

*Co-authored work: The signatures of all authors are required. Each signature attests to the originality of the entire piece of written work in its final form.

Print form

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of Originality

This sheet must be signed and enclosed with every piece of written work submitted at ETH.

I hereby declare that the written work I have submitted entitled

Refugee Modelling

is original work which I alone have authored and which is written in my own words.*

Author(s)

Last name	First name
Pietrasik	Lukasz

Supervising lecturer

Last name	First name
Sanders	Lloyd

With the signature I declare that I have been informed regarding normal academic citation rules and that I have read and understood the information on 'Citation etiquette' (http://www.ethz.ch/students/exams/plagiarism_s_en.pdf). The citation conventions usual to the discipline in question here have been respected.

The above written work may be tested electronically for plagiarism.

Zürich 12th Dec 2016
Place and date

Lukasz Pietrasik
Signature

*Co-authored work: The signatures of all authors are required. Each signature attests to the originality of the entire piece of written work in its final form.

Print form

ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of Originality

This sheet must be signed and enclosed with every piece of written work submitted at ETH.

I hereby declare that the written work I have submitted entitled

Refugee Modelling

is original work which I alone have authored and which is written in my own words.*

Author(s)

Last name	First name
Schumacher	Sylvia

Supervising lecturer

Last name	First name
Sanders	Lloyd

With the signature I declare that I have been informed regarding normal academic citation rules and that I have read and understood the information on 'Citation etiquette' (http://www.ethz.ch/students/exams/plagiarism_s_en.pdf). The citation conventions usual to the discipline in question here have been respected.

The above written work may be tested electronically for plagiarism.

Zürich, 12th December 2017 — S. Schumacher
Place and date — Signature

*Co-authored work: The signatures of all authors are required. Each signature attests to the originality of the entire piece of written work in its final form.

Print form

Contents

1 Abstract **1**

2 Individual contributions **1**

3 Introduction and Motivations **2**

4 Description of the Model **3**
4.1 Stouffer's theory of migration 1960 3
4.2 Adapting Stouffer's theory of migration to our problem setting 3

5 Implementation **5**
5.1 Basic Data Investigation 5
5.1.1 Input data 5
5.1.2 Model fitting 7

6 Simulation Results and Discussion **7**
6.1 Correlation between total number of Syrian refugees and deaths in Syria . . 7
6.2 Model estimation and model fit 8
6.2.1 Model estimation 8
6.2.2 Assessing model fit 8
6.2.3 Model coefficients 10
6.3 Out of sample prediction for 2015 10
6.4 Model improvements 13

7 Summary and Outlook **14**

8 References **16**

1 Abstract

In 2015 European governments have struggled to accommodate the high number of incoming refugees from Syria due to the military conflict in the region. This unexpected number of refugees have created the European refugee crisis started in 2015. This paper aims to investigate whether the 2015 European refugee crisis and their spread among Europe could be predicted with the use of Stouffer's Theory of Mobility before it has happened. In order to find the answer, the correlation between the number of Syrian deaths in war and the total number of refugees has been made giving the high value of the corresponding Pearson correlation coefficient ($r = 0.85$). Furthermore, the modifications to the Stouffer's Theory of Mobility formula has been made to calculate the number of refugees in each European country given the following values: the number of deaths in Syria, the GDP of the given country, the calculated sum of the GDP of the countries on the way to the destination, and the number of Syrian refugees already in the country, where each indicator reflects information from the previous year. The training of the parameters on the historical data for years 2010-2014 have resulted in satisfactory values. It has been proven, that the model was able to predict the spread of European refugees in 2015 with the final equation. However, the model only partly succeeds in predicting the overall migration flow of Syria to Europe. The particularly interesting finding was the correlation between the number of refugees coming to the given country and the number of Syrian refugees already in the country being proportional instead of inversely proportional as expected from the initial model. The obtained equation has predicted the number of refugees for every European country, giving the coefficient of determination of $R^2 = 0.873$ and $R^2_{adjusted} = 0.870$, proving a very good fit of the model to the real word data.

2 Individual contributions

Team members:

Lukasz Pietrasik has implemented the first iteration of the Stouffer's Theory of Mobility into MATLAB and contributed to the final report.

Sylvia Schumacher has gathered and merged empirical input data for the research project, visualized the underlying information and correlations using R and contributed to the final report.

Wojciech Witon has mastered the model implementation in MATLAB and plotted the final results of the linear regression equation.

3 Introduction and Motivations

"We are facing the biggest refugee and displacement crisis of our time. Above all, this is not just a crisis of numbers; it is also a crisis of solidarity."

Ban Ki Moon, United Nations Secretary General [2]

The number of refugees has increased significantly in recent years and has reached 21.3 million in 2015 [10]. The unstable political situation, ongoing armed conflict and persecution influence refugees' decision-making process of leaving their country. Almost one-fourth (4.9 million) of all refugees in 2015 originate from the Syrian Arab Republic. A lot of people looking for an asylum have chosen Europe as a stable place to live thanks to its decent location and economic conditions [10]. The uncontrollable and unforeseen movement became a political challenge for European countries and finally lead to the European refugee crisis in 2015. Is there any theory that would aid the governments by reliable prediction of the number of refugees applying for asylum in the given country?

Stouffer's Theory of Mobility [4] is chosen to simulate reliable predictions, which is motivated in the Appendix in chapter 8. Stouffer's Theory of Mobility was originally developed in 1940. In contrast to the Gravity model, Stouffer's model does not assume a direct relationship between migration and distance. Instead, a concept of intervening opportunities is introduced. For this reason, the model is also called an "Intervening Opportunities model". The main idea behind the model is, that migration flows in space are spread inversely proportional to the number of opportunities between the place of origin and destination [11]. Additionally, the migration flow is proportional to the number of opportunities in the country of destination (X_1).
Stouffer extended his basic model in 1960 by adding the concept of competing migrants into his model [13]. The idea behind the modified version is, that migration is inversely proportional to the number of migrants in the place of destination since they are competing for economic needs such as labor, real estate, etc. [12]. The modified version of Stouffer's theory can be found in chapter 4.1.

Additional to Stouffer, who predicted interstate migration with his model, Galle and Taeuber (1966) proved good performance of Stouffer's migration theory when applying the competing migrant model to metropolitan migration flows from 1955-1960 [6].

Hence, the main research question stated for this project is as follows:
Could the 2015 European refugee crisis and the refugee spread among Europe have been predicted with the use of Stouffer's Theory of Mobility before it has happened? This research question implies need to accurately predict the overall flow of refugees coming to Europe in 2015 as well as their spread among respective European countries.

The goal was to examine whether an intuitively suitable simulation model could accur-

ately reflect Syrian refugee flows during the European refugee crisis in 2015.

Understanding which factors are crucial for the competitive attractiveness of European countries, the model yields different conceivable generalization approaches: first, under the assumption that factors influencing the migration decision-making process stay the same, the model can be used to predict migration flows of years further in the future. Second, the simulation could help in the development of future migration and boarder policies by understanding and orchestrating modifiable factors of refugee movement. Furthermore, the same model could be applied to estimate the number of asylum seekers in countries that lack empirical data of historic refugee movement flows.

To analyze the research question, the locations of Syrian first-time asylum applicants were examined. The underlying empirical data consists of annual movement flows from 2000 to 2015.

4 Description of the Model

4.1 Stouffer's theory of migration 1960

According to [12], the number of migrants between two cities 1 and 2 is denoted as Y, which can be formulated by:

$$Y = \frac{K * (X_0 * X_1)^A}{X_B^B * X_C^C} \tag{1}$$

In equation 1 the number of migrants between two cities is proportional to the number of all migrants in the city of origin (X_0) and the number of opportunities in the city of destination (X_1). Furthermore, the model assumes, that this number is inversely proportional to the number of opportunities intervening between the two cities of examination (X_B) and migrants competing in the city of destination (X_C). K denotes a scaling constant and is determined empirically as well as the parameters A, B, and C.

4.2 Adapting Stouffer's theory of migration to our problem setting

The main assumption of the model used is that the proportion of refugees' among peaceful countries can be predicted by the modified Stouffer's Theory of Mobility. According to the original theory, the proportion of all migrants from the place of origin moving to the place of destination is defined as:

$$\frac{X_1}{X_B * X_C}. \tag{2}$$

Multiplying this proportion by the number of all migrants from the place of origin (X_0) will thus be proportional to the number of migrants moving between two places of interest.

However, in our problem setting, the initial decision about leaving the country is different than in Stouffer's original case of migration between cities. It is not driven by competition in the corresponding region but by the political situation, and state of war. To reflect this accordingly, the model predicts the total number of people deciding to leave the country as being proportional to the 'violence rate' in the place of origin, Syria. The predictions for a total number of people leaving the country are modeled by the following formula:

$$N = a * V_r^R, \tag{3}$$

where N denotes the number of people leaving the country, V_r is the measure of violence rate in the country. Thus the assumption, that with increasing violence rate, the decision making of leaving the country increases exponentially. The motivation behind this approach is, that at a low level of violence, people would be resistant against leaving their country which implies giving up family, friends, and households, but with increasing risk, their willingness to leave increases disproportionately. The values a and R are constants determined empirically.

By substituting the X_0 from Equation 3 by the N from the Equation 2 and simplifying it, the final model formula is determined:

$$Y = K' * \frac{V_r^{R'} * X_1^A}{X_B^B * X_C^C} \tag{4}$$

where values $V_r\ X_1,\ X_B,\ X_C$ are as described above with $K' = K * a^A$ and $R' = R * A$.

Whereas [11] and [6] each inspect movements in one specified time interval, this work gathers annual panel data from 2010 to 2015.

Since the research tries to answer the question, whether it was possible to forecast the European refugee crisis and Syrians spread among Europe, the model was trained predictors of the foregoing year. This reflects, that in order to predict the refugee crisis in 2015 before it has happened, politicians only had information about economic indicators of the year 2014. As such, the number of migrants moving from Syria to the country of destination (c) in year t ($Y_{t,c}$) is modeled as:

$$Y_{t,c} = \frac{V_{r,(t-1)}^R * X_{1,c,(t-1)}^A}{X_{B,c,(t-1)}^B * X_{C,c,(t-1)}^C} * K. \tag{5}$$

5 Implementation

5.1 Basic Data Investigation

5.1.1 Input data

In order to train the model on real-world empirical data, the **number of Syrians first time asylum applications** in the respective European country (Y) was used. The data was provided by the UN refugee Agency (UNHCR) [9]. However, it is important to realize that due to Schengen zone the people are able to change the country after obtaining the refugee status. Nonetheless, it seems reasonable, that a large proportion of refugees do not move further after applying for asylum in one country.

The refugee migration model requires the measure of violence input in order to estimate the total number of refugees in a given year. Following the statistical analysis the number of Syrian deaths in the war was used as the indicator of the **measure of violence** (V_r) in the country of origin. For that purpose the empirical time series data originates from the "Uppsala Conflict Data Program" (UCDP) [1] was obtained and studied. The total number of deaths was determined to indicate a push factor from the country of origin and was further used to predict the total number of people leaving the country, which corresponds to X_0 in the original model.

The **number of opportunities** ($X_{1,c,(t)}$) at time (t) is modeled by the overall GDP of the the European country c. In contrast to the GDP per persona, the overall GDP yields the advantage, that it implicitly involves information about the population size of a given country. Hence the reasoning that more dense populated countries might attract a higher number of refugees than smaller ones. The data originates from Knoema [8].

Figure 1 shows the average of the total Syrian refugee flow divided by the GDP of the respective European country from 2012 to 2015. The figure implies, that GDP is not the only factor, moving the spread of refugees among Europe. As can be seen, in particular, Cyprus, Bulgaria, Malta, Sweden, Hungary, Greece, and Germany do have very high Syrian refugee migration compared to their GDP. While Cyprus and Malta are very small countries, that might be negligible our remaining model indicators should be able to explain the variation in the countries of bigger size. Portugal, Estonia, Slovakia, and Poland are at the bottom of the figure, implying that they have comparatively few refugees. Intuitively, this is in alignment with the countries' policy with respect to foreigners.

The **intervening opportunities** are calculated by summing all the GDPs of the countries that refugees have to cross in order to get to a given country B, can be calculated with:

$$X_{B,c_i} = \sum_{c_j \in C : c_j \neq c_i} GDP_{c_j} * \mathbb{1}_{c_j} \tag{6}$$

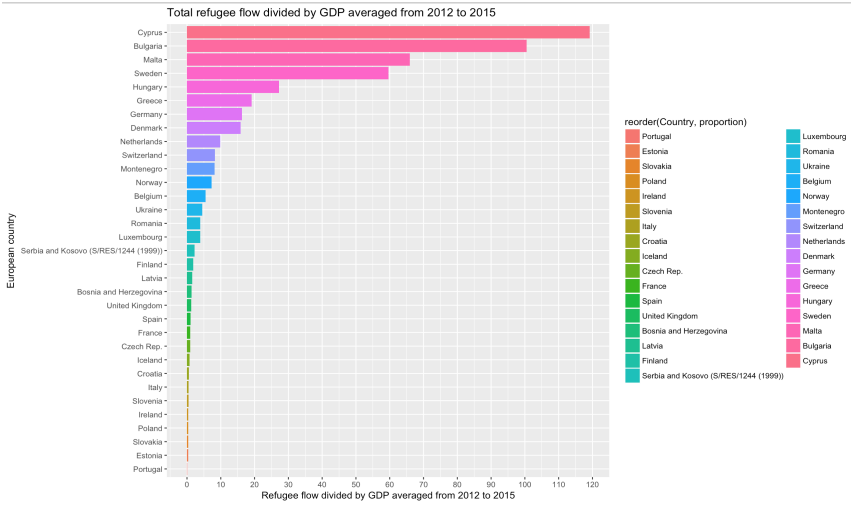

Figure 1: Total refugee flows divided by averaged GDP from 2012 to 2015

In order to compute these values the 'route matrix' has been determined that contains 0 or 1, where:

$$\mathbb{1}_{c_j} = \begin{cases} 1 & \text{if country j is crossed on the way from Syria to the country of destination} \\ 0 & \text{else} \end{cases}$$

and C is defined as the set of all countries of destination.

The values of route matrix have been manually determined by investigating the shortest route measured in traveling time computed by Google maps on 18 November [7]. All routes have been measured from Turkey as a starting point and if the times of multiple suggested routes were similar (with differences smaller than one hour) the total distance was minimized.

The last indicator for the modeling equation is the number of Syrians in the country of destination as a measure of the **competing migrants** (X_C). Since there are no information of interest available, the approximation of this indicator was calculated by summing all Syrians who have applied for asylum in a given European country from 2000 until the time of interest (t-1):

$$X_{c,(t)} = \sum_{t_i=2000}^{t-1} SyrianRefugees_{t_i,c} = \sum_{t_i=2000}^{t-1} Y_{t_i,c}. \tag{7}$$

This approach can be justified by the fact that the number of asylum seekers from Syria

before 2000 was very low compared to the numbers in the subsequent years so that the competition of Syrians entering Europe before 2000 is negligible.

5.1.2 Model fitting

For estimation purposes of the factors K, R, A, B and C, the model was linearized by taking the logarithms of the left and right hand side of the equation.

$$\log(Y_{t,c}) = \log(K)+R\cdot\log(V_{r,(t-1)})+A\cdot\log(X_{1,c,(t-1)}))-B\cdot\log(X_{B,c,(t-1)})-C\cdot\log(X_{C,c,(t-1)}) \tag{8}$$

The regression was made based on data from the years 2010 to 2014 leaving the full data of 2015 for the evaluation of the prediction accuracy of the results. For ease of interpretation, the model parameters K, A, B and C as well as R do not vary over time, which can be motivated by the assumption, that patterns of the migration decision-making process do not change significantly over time.

6 Simulation Results and Discussion

6.1 Correlation between total number of Syrian refugees and deaths in Syria

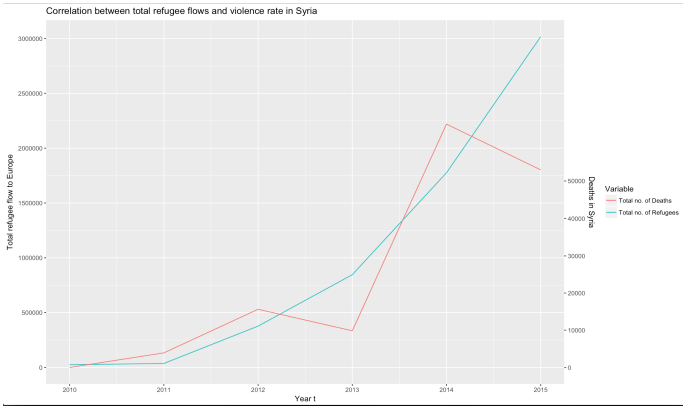

Figure 2: Correlation between total number of Syrian refugees and deaths in Syria

The correlation between the number of Syrian deaths and the total number of refugees in a given year is comparatively high with a corresponding Pearson correlation coefficient

($r = 0.85$). The results are showed in detail in Figure 2. It can be seen that there exists a strongly increasing time trend in both variables of investigation: the number of deaths increased from almost 0 in 2010 to almost 3 million in 2015. The total number of deaths can be seen as a push factor from the country of origin and is further used to predict the total number of people leaving the country, which corresponds to X_0 in the original model.

6.2 Model estimation and model fit

6.2.1 Model estimation

The proposed equation 8 was fitted with the use of MATLAB for the years 2010 to 2014 resulting in the values presented in Table 1.

Table 1: Summary output of model fit based on training data from 2010-2014

	Estimate	SE	T Statistics	p Value
Scaling factor K	-0.378	0.112	-3.360	0.001
Total deaths	0.160	0.022	7.296	0.000
GDP	0.221	0.060	3.714	0.000
GDP between	0.201	0.045	4.484	0.000
Previous number of refugees	-0.878	0.039	-22.683	0.000

Substituting the values into the initial equation, with K obtained by taking exponential of the Intercept, the following formula has been obtained:

$$Y = 0.6852 * \frac{X_0^{0.160} * X_1^{0.221} * X_C^{0.878}}{X_B^{0.201}} \tag{9}$$

Table 1 shows that all of the chosen indicators are highly significant, with p-values below the 1%-level. Thus, the model will lose explanatory power, if the model indicators were reduced.

6.2.2 Assessing model fit

The resulting coefficient of determination is $R^2 = 0.873$ with $R^2_{adjusted} = 0.870$. The high values show a well-fitted model for the problem setting.

The scatterplot in figure 3 shows the relation between the historic number of refugees in a given country versus the prediction values for the chosen training years (2010 to 2014). The ideal model fit would contain all data on one, straight line at 45° angle that passes through zero. The data points are clearly gathered around the perfect line and do barely

show significant out-layers. The model identifies larger fitted values more precisely than smaller values since the variation from the diagonal becomes smaller with increasing refugee flows. This means that for countries with larger predicted refugee flows such as Germany, the prediction is of higher precision. However, for countries with smaller refugee flows, such as Malta or Estonia, we can not precisely predict Syrian migration.

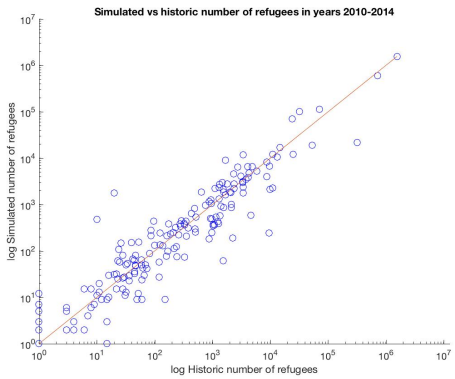

Figure 3: Scatterplot (in sample) of true vs. fitted values from 2010 to 2015

In order to inspect whether regression assumptions are met, the Tukey Anscombe plot was prepared in Figure 7 in the Appendix. Figure 7 also confirms, that the variance of residuals $Var(E_i)$ massively decreases for larger fitted values.

The plotted residuals on Figure 4 shows that the final model overestimates the number

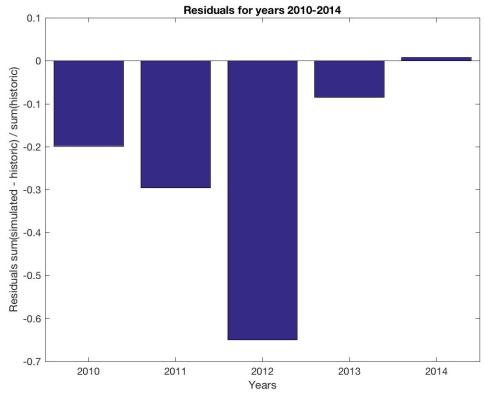

Figure 4: Yearly distribution of the residuals normalized by the true migration flow values

of refugees in Europe for the time period of 2010 to 2013. The lowest value of the residual occurs in 2012 and is equal to -0.65. However, for 2014 the data is underestimated with the residual value being close to 0. Starting from 2013 the number of refugees from Syria has started increasing, with the most significant number in the year 2014. These values are more valuable in the regression model and it can be seen from Figure 4 that the residuals indicate that the prediction is closer than for other years with the best fit for 2014.

6.2.3 Model coefficients

As expected from Stouffer's Theory of Mobility, the positive estimates for the parameters K, R, A and B were received. Contradictory from Stouffer's Theory of Mobility [12] the parameter C is positive which means that the value X_C, reflecting competing migrants in the original model, is in the numerator instead of the denominator [6]. One explanation for this would be, that refugees differ from economic migrants as mainly considered in [12] and [6]: instead of competing at the country of destination, they help each other which creates an attracting and not repulsing force among refugee groups. Another interpretation is, that a higher cumulative number of refugees at a given country could reflect the immigration law and border policy. We did not incorporate any indicator of the countries' friendliness towards refugees and X_C might be highly correlated with this omitted variable, thus catching up the explanatory power of this important information: countries that are more welcoming to refugees attract higher migration flows.

6.3 Out of sample prediction for 2015

The obtained model equation 9 was used to estimate the number of refugees coming to each European country in 2015. Since the model was trained only on data from 2010 to 2014, the forecast for 2015 was an out of sample prediction. The model was also applied to forecasting 2016 migration flows based on indicators from 2015 for future evaluations. The results can be found in the Appendix 8.

After running the prediction for 2015 the proposed research question can be evaluated: Could the 2015 European refugee crisis and the refugee spread among Europe have been predicted with the use of Stouffer's Theory of Mobility, before it has happened?

Overall historic migration flows and estimated migration flows for 2015 were compared. Overall migration flows were overestimated by 70% as can be seen in the first row of table 2.

Figure 5 explores the differences of the true values for 2015 and the predicted values on a country-level perspective. It can be seen, that the model overestimates the migration flows in 2015 for most countries. This is in line with the overall overestimated migration flow. The reason for this overestimation might be found in Figure 2. The number of deaths

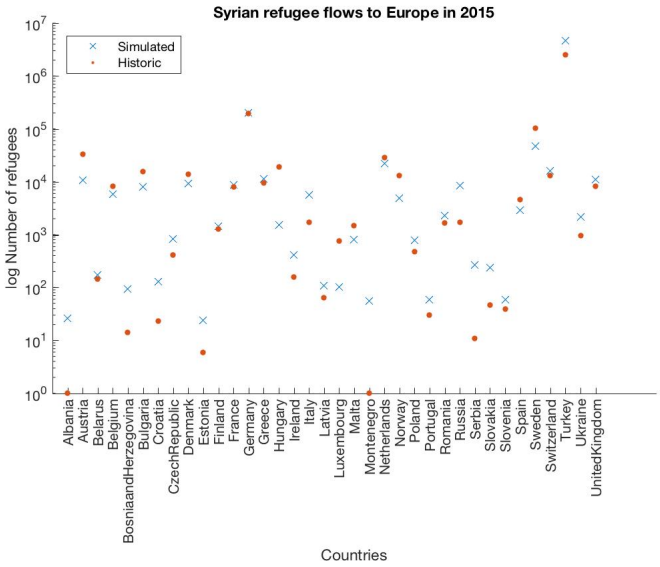

Figure 5: Out of sample prediction and true historic values for Syrian refugee flows to Europe in 2015

in Syria reached a maximal level in 2014 with a total value of 65.281. Since the indicator value from 2014 is used for the prediction of 2015 and has a positive regression coefficient of 0.16, the high number of deaths leads to the overestimation of the overall migration flow. Combining the information from plot 9 and table 2 the conclusion can be made, that the model was not capable of capturing the overall migration flows to Europe in 2015.

A more detailed analysis shows, that estimates for Turkey deviate by almost 90% from the truth. The main reason for this might be, that Turkey was used as the starting country of the proposed route matrix. As such, there are no intervening opportunities that are crossed on the way to Turkey and the model does not properly distribute these refugees

Table 2: Comparison of estimated overall migration flows 2015

	Historic Value	Predicted Value	Deviation
Overall migration flow	2.989.613	5.089.830	+0.703
Overall migration excluding Turkey	485.767	386.485	-0.204
Turkey	2.503.846	4.703.345	+0.878

to other countries. In order to further analyze the prediction of refugees among European countries, Turkey was excluded from further analysis. Further research has to be done to solve the problem of distributing refugees from the start of the resource matrix. Attempts to solve this problem are pointed out in chapter 6.4.

However, as can be seen in the second row of table 2, when excluding Turkey from the overall migration flow of Europe, the deviation of the true and estimated values can be reduced and migration flows are underestimated by 20.4%, which is an accurate prediction performance.

In the following, the model performance with respect to its ability of predicting the spread of refugees among Europe shall be assessed:

The coefficient of identity yields a high value of $R^2 = 0.908$. From a naive point of view, it can be said, that 90.8% of the overall variation are explained by the model implying a relatively high explanatory power of the model.

Additionally, the overall goodness of fit is assessed with Pearson's χ^2 test for categorial variables. The predicted and true values are clustered in countries, allowing for a χ^2 test. The test measures the discrepancy between the true historic distribution and the values estimated by the proposed model [14]. The test statistics follows a χ^2-distribution with $(k-1) = 34$ degrees of freedom, where k is the number of categories of the observed values. The null hypothesis that the historic and predicted values are independent of each other can be successfully rejected with a p-value of $< 2.2e-16$ of the corresponding test statistics $\chi^2 = 55372$. The root mean squared error calculated by

$$RMSE = \sqrt{\frac{\sum_{i \in C} (\hat{Y}_i - Y_i)^2}{|C|}}, \tag{10}$$

with C as the set of all observed countries yields $RMSE = 10971.82$ for the predictions in 2015. The interpretation of the RMSE would be, that the standard deviation of the unexplained variance on average is 10.971 refugees per country.

In Figure 6 can be observed, that very small migration flows such as in Albania and Montenegro are strongly overestimated compared to their real value. The model introduced will never estimate values of zero due to its additive components. Since these refugee flows only display small proportions of the overall migration, the importance of estimating these values very precisely is not given.

In Figure 6 it can also be seen, that the model underestimates migration flows for eleven countries: Austria, Belgium, Bulgaria, Denmark, Hungary, Luxembourg, Malta, Netherlands, Norway, Spain and Sweden. These are only eleven representatives where predicted values are underestimated, though leading to the overall underestimation of migration flows in Europe. Most of these countries are nowadays known for having welcomed many refugees during the European refugee crisis. The model does not sufficiently succeed in forecasting the dense distribution of refugees among these countries. With exception from Spain, these

countries can also be found on the upper half of figure 1. This implies, that the ratio of refugee flows divided by the countries' GDP is also in the upper ranges. Although GDP and historic refugee flows are two of the main model inputs, the model cannot predict the proportion of refugees entering these countries being higher compared to other countries. As can be seen in figure 6, the model performs best for the countries Belarus, Belgium, Denmark, Finland, France, Germany, Greece, Netherlands, Romania and Switzerland. These countries do mainly have slightly above average ranks in figure 1, implying, that they do have an average migration flow compared to their GDP. It can be concluded, that the model provides better predictions for these countries compared to countries with a more unbalanced proportion of GDP and migration flow. The top ten countries have a tendency to be geographically spread between Western and Northern Europe.

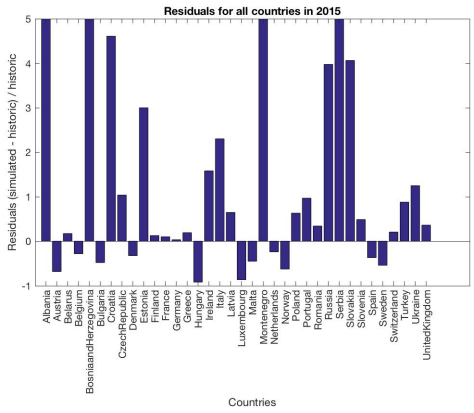

Figure 6: Distribution of the residuals among countries for 2015

It is worth to mention Bosnia and Herzegovina, Croatia, Estonia, Russia, Serbia and Slovakia are merely identified correctly. It can be recognized, that these countries are mainly in the South-East of Europe and thus close to Turkey. Since the route matrix for the intervening opportunities originated from Turkey, there are only a few intervening opportunities on the way crossed by these migration flows. As such, the value in the denominator of our estimation equation stays small, resulting in higher predicted refugee flows than finally realized. Further simulations with more sophisticated route matrices should be run to estimate results more precisely as discussed in section 6.4.

6.4 Model improvements

Our measure of competing migrants is the cumulative sum of Syrians entering from 2000 to the year previous of the investigation as described in equation 7. However, we do not take into account asylum seekers from other nations. These groups might affect Syrians

since they compete with respect to real estate, labor or welfare.

To compute the intervening opportunities between two places, we do sum up opportunities on the shortest route between origin and destination so far. The route matrix should be more sophisticated by introducing multiple starting points and not only Turkey as in the current version. Furthermore, only the fastest route from the origin to destination is calculated so far. This procedure can be extended by using multiple different routes and assigning proper weights to them.

Different indicators could contribute in order to forecast the distribution among European countries more precisely: a measure of the countries respective border policy or their welfare plans for asylum seekers might be omitted variables with high explanatory power. However, it is difficult, to measure these aspects in a quantitative form and incorporate it into the model.

We use time series data of the total refugee flows coming from Syria to a respective country as input (in form of the cumulative sum up to year $(t-1)$) and output variable. This leads to the presence of autocorrelation in the error terms. Statistical tests are rejected too often since the t-statistics are unreliable. Further research could try to work with methods for Time Series Analysis, which model cross-time correlation structure. Additionally, the proposed model parameters do not vary over time. This means, that the respective indicators do always have the same impact on the prediction of refugee movement over time. Loosen this restriction could yield further insights about changes in migration patterns.

An alternative approach would have been an agent-based modeling approach since the decision-making process of refugees also depends on demographic and geographical patterns. This approach might be better in predicting overall refugee flows by explicitly modeling personal patterns. However, it is difficult to find reliable data where such a model can be build upon. In the Appendix in section 8, the proposed model is also compared to the commonly used Gravity model.

To increase internal validity, more detailed data containing shorter time intervals (on a quarterly or monthly basis), allows for more detailed predictions and takes seasons into account. To check external validity and robustness, the same approach could be applied to other conflicts and countries.

7 Summary and Outlook

A modified version of Stouffer's Theory of Mobility was applied to simulate Syrian migration flows. The equation parameters were obtained using the linear regression of the logarithmic model resulting in the following equation:

$$Y = 0.6852 * \frac{X_0^{0.160} * X_1^{0.221} * X_C^{0.878}}{X_B^{0.201}}, \tag{11}$$

where Y reflects the number of refugees in given country, X_0 is the number of deaths in Syria in the previous year, X_1 is the GDP of the given country, X_B is the sum of the GDP of the countries on the way to the destination and X_C reflects the number of Syrian refugees already in the country in the previous year.

An interesting finding is the presence of the X_C value in the numerator instead of the denominator as proposed in the original model. This can be led back to refugees who are not competing but might rather help each other creating an attractive force for others to come.

The above equation was used to predict the number of refugees for the year 2015 for every European country and to answer the research question, whether it was possible to predict the European refugee crisis in advance. It can be stated, that the strong increase in the total number of deaths clearly indicated an increased migration flow from Syria to Europe before the European refugee crisis has happened which our model correctly recognizes. Although the model is able to predict a strong increase in migration flows, the exact extent of this increase is not identified precisely for 2015.

To account for the overestimation of the migration flow and in order to be able to compare the model ability to predicting the spread of refugees among Europe, Turkey was excluded from further analysis. Thus, the overall deviation could be reduced to 20.4% for 2015. The χ^2 Goodness of Fit test with a p-value of $< 2.2e-16$ showed clear dependencies between the predicted and the true historic values as well as the coefficient of identity with a value of 0.908. The model predicts well the spread of refugees among countries with average migration flows compared to their GDP. The model does not accurately recognize most countries, that are strongly welcoming refugees and exhibit very dense refugee flows compared to their GDP. Less than one third of all countries was underestimated, though leading to an overall underestimation of migration flows in Europe excluding Turkey. The ten countries with the best predictive values compared to their overall migration flow in 2015 have a tendency to be geographically spread between Western and Northern Europe. Considering relative prediction errors from a country-level perspective, countries in the South-East of Europe and thus close to Turkey perform worst. Summarizing these results, a more sophisticated route matrix should be evaluated.

We conclude, that is was very difficult to predict the European refugee crisis in 2015 before it has happened. This was also revealed in practice, where the migration flows were not foreseen properly. Our adapted version of Stouffer's Theory of Mobility needs further improvements to predict the refugee crisis properly.

In the future, it will not only be necessary to predict refugee flows more concisely, but also put effort in avoiding them. The research discovered the high correlation between the number of Syrian deaths and the total number of refugees ($r = 0.85$). This result shows possible incentive for European governments to prioritize the actions that aim to stabilize the region and protect the Syrian population in the future.

8 References

References

[1] Ucdp - uppsala conflict data program. Ucdp.uu.se [Accessed: 2016-11-14].

[2] Refugee crisis about solidarity, not just numbers, secretary-general says at event on global displacement challenge, 2016. http://www.un.org/press/en/2016/sgsm17670.doc.htm [Accessed: 2016-12-03].

[3] Yongwan Chun. Modeling network autocorrelation within migration flows by eigenvector spatial filtering. *Journal of Geographical Systems*, 10(4):317–344, 2008.

[4] Gordon F De Jong and Robert W Gardner. *Migration decision making: multidisciplinary approaches to microlevel studies in developed and developing countries*. Elsevier, 2013.

[5] Kayo Fujimoto, Chih-Ping Chou, and Thomas W Valente. The network autocorrelation model using two-mode data: Affiliation exposure and potential bias in the autocorrelation parameter. *Social networks*, 33(3):231–243, 2011.

[6] Omer R Galle and Karl E Taeuber. Metropolitan migration and intervening opportunities. *American Sociological Review*, pages 5–13, 1966.

[7] Google Inc, 2016. https://www.google.ch/maps [Accessed: 2016-12-17].

[8] Knoema. Imf world economic outlook (weo). https://knoema.com/IMFWEO2016Apr/imf-world-economic-outlook-weo-april-2016 [Accessed: 2016-11-17].

[9] Popstats.unhcr.org. Unhcr population statistics - data - time series. http://popstats.unhcr.org/en/time_series [Accessed: 2016-11-11].

[10] United Refugees. Unhcr global trends 2015, 2016. http://www.unhcr.org/statistics/unhcrstats/576408cd7/unhcr-global-trends-2015.html [Accessed: 2016-12-3].

[11] Samuel A Stouffer. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 5(6):845–867, 1940.

[12] Samuel A Stouffer. Intervening opportunities and competing migrants. *Journal of regional science*, 2(1):1–26, 1960.

[13] Walter J Wadycki. Stouffer's model of migration: A comparison of interstate and metropolitan flows. *Demography*, 12(1):121–128, 1975.

[14] J Watkins. Introduction to the science of statistics. 2016. http://math.arizona.edu/~jwatkins/U_gof.pdf [Accessed: 2016-12-10].

Appendix

Checking regression model assumptions

The Tukey-Anscombe plot shows the residuals r_i vs. fitted values $\hat{y}_i$ and verifies whether $E[E_i] = 0$, i.e. whether the model makes unbiased predictions. As can be seen from Figure 7, the residuals approximately scatter around 0 and thus we met the assumption of unbiased predictions ($E[E_i]0 =$).

As mentioned in chapter 6.3 the variance of residuals $Var(E_i)$ decreases for larger fitted values. This means, that the regression assumption $Var(E_i) = \sigma_E^2$ is not met. However, this assumption is less severe and since we are mainly interested to estimate the main part of the spread of refugees correctly, i.e. to estimated correct large fitted values, we neglect this GLM assumption violation.

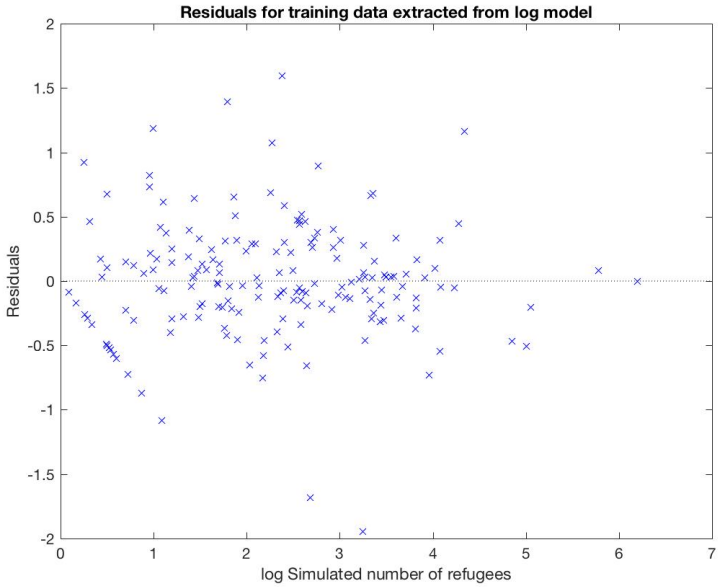

Figure 7: Tucey Anscombe plot for model trained on data from 2010 to 2014

Simple OLS approach

Figure 8 illustrates a simple linear regression of the overall refugee flows from Syria to Europe. The regression line is trained solely on the historic data of incoming refugees from 2010 to 2014. As can be seen, the trend in this regression line is not able to explain the European refugee crisis in 2015. According to the regression line, we estimate a value of

1.905.527 incoming Syrian refugees which deviates 36.8% from the true value of 3.016.635. A suitable model must be able to significantly reduce this deviation.

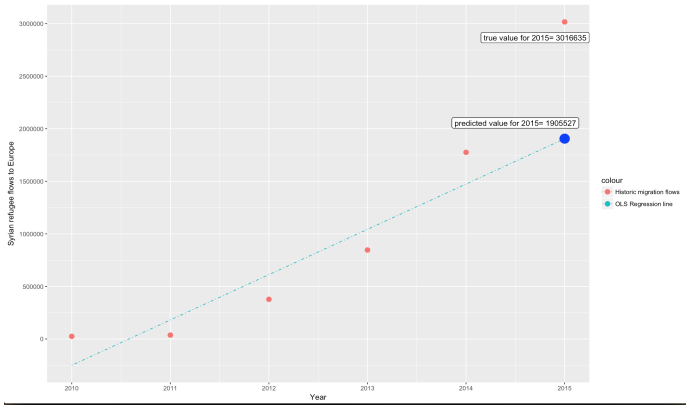

Figure 8: OLS approach to explain the Syrian migration flow from historic data

Forecast 2015

The following figure compares the overall residuals of the 2015 forecast with the residuals of former forecasts:

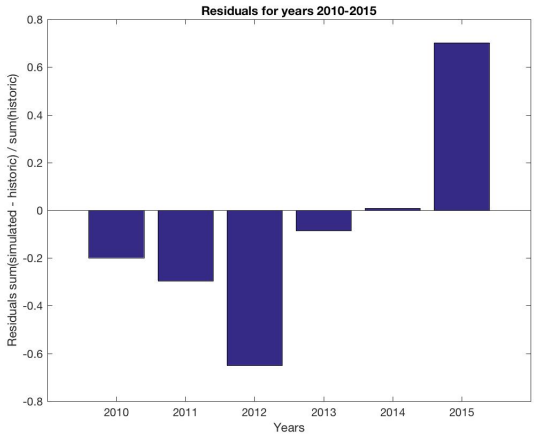

Figure 9: Yearly Distribution of the residuals including 2015

Forecast 2016

We estimate the following migrations flows from Syria for 2016:

Table 3: Refugee number simulation in 2016

Albania	25
Austria	28627
Belarus	248
Belgium	9350
Bosnia and Herzegovina	96
Bulgaria	13681
Croatia	138
Czech Republic	1080
Denmark	14412
Estonia	27
Finland	2126
France	14307
Germany	321811
Greece	15802
Hungary	11279
Ireland	484
Italy	7040
Latvia	136
Luxembourg	439
Malta	1194
Montenegro	55
Netherlands	33745
Norway	10513
Poland	1161
Portugal	77
Romania	3246
Russia	9273
Serbia	256
Slovakia	261
Slovenia	83
Spain	5700
Sweden	80713
Switzerland	22267
Turkey	8041376
Ukraine	2482
United Kingdom	15998

Model choice

According to De Jong and Gardner (2013) various theoretical and empirical migration modeling approaches do have two dimensions: in the first one population aggregates or areas and their explanatory patterns driving the decision making process of leaving a country are of interest. In this literature, demographic factors such as age, gender, race and home-ownership or migration between regions are examined with an attempt to find causal relations with migration. The second dimension considers distance and volume of movement flows between areas. Among these models, the most common representative is the gravity model originated from physical science. According to the gravity model, migration flows are proportional to population sizes in the place of origin and destination and inversely proportional to the distance between them [4].

Compared to other migration models, Stouffer's theory of Mobility has the advantage, that it explicitly models dependencies among areas by their intervening opportunities. The problem setting of this research is to understand the decision making process of refugees. Since Syrian refugees mainly come for political and not for economical reasons, distance might not accurately reflect their spread in space. Nowadays, it is comparatively easy to travel also long distances the deviations in distance from Syria to any European country is on a moderate level anyway. What might matter more, are the opportunities in the respective countries on the way they cross.

Thus, compared to a gravity model, our model also takes network autocorrelation into account. Network correlation means, that migration flows are not independent of each other. In our case, the decision making process of leaving Syria might be correlated with the social and environmental context in Syria and the choice of the country of destination might depend on the choice of other peers [5]. For this purpose, dealing with Network autocorrelation is crucial. We model the dependence among migration flows by explicitly incorporating intervening opportunities (X_B) and competing refugees (X_C) into our model. Both variables reflect interactions in between migration flows.

In the gravity model, only the overall distance from the place of origin to the place of destination is taken into account. This can lead to a misspecification of the model equation and the estimation may become unreliable.

Chun (2008) shows, that an alternative modeling approach to deal with this autocorrelation would be Poisson regression with eigenvector spatial filtering [3].