A Classification and Data Visualization Tool applied to Human Migration Analysis

David Dominguez, Pablo Soria, Mario González,
Francisco B. Rodríguez and Ángel Sánchez

May 28, 2019

A Classification and Data Visualization Tool applied to Human Migration Analysis

1st David Dominguez
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain
david.dominguez@uam.es

2nd Pablo Soria
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain

3rd Mario González
SI^2 Lab
Universidad de las Américas
Quito, Ecuador
mario.gonzalez.rodriguez@udla.edu.ec

4th Francisco B. Rodríguez
Escuela Politécnica Superior
Universidad Autónoma de Madrid
28049 Madrid, Spain
f.rodriguez@uam.es

5th Ángel Sánchez
ETSII
Universidad Rey Juan Carlos
28933 Móstoles (Madrid), Spain
angel.sanchez@urjc.es

***Abstract*—Nowadays, in a highly-globalized world, the understanding of causes and consequences involved in the migration phenomena, and also the prediction of migration flows are important for development of national public policies and for urban resource planning. The high complexity of human im/emigration movements can not only be explained by economic causes but rather by the interaction among multiple additional factors (demographic, social, linguistic, among others). The application of Machine Learning techniques and Data Visualization models on high volumes of raw data from countries can provide good insight to understand how indicators from countries are related to migration causes, and also to make visible the migration flows between the sending and receiving countries. This paper describes a tool which includes supervised classification and visualization methods to analyze country indicators and aim to discover the connections among these attributes and the migration movements.**

***Index Terms*—Data Visualization, Country Classification, Socio-Economic Indicators, Country Indicator Histogram, Migration Map, Multilayer Perceptron**

I. Introduction

It is estimated that in 2015 the number of international migrants was nearly 250 million people. This number has been tripled in the last 50 years, and so the human migrations have become a complex global challenge. Migration is a kind of human mobility, where the displacements involve people moving from one to another geographic region and/or country, and changing their home locations. People who leave their country are called emigrants and those who move into another different country are called immigrants. Although in some cases people have the choice to move, it is in most cases that they are forced to migrate due to pushing factors (e.g. wars, lack of work, bad economical conditions or natural disasters, among others). It is currently difficult to find reliable information on migration flows [1]. This information, as well as the accurate prediction of future migrations [2], is crucial to take correct decisions to mitigate the migration-related effects, which have a crucial influence in the development of public socio-economical policies of receiver countries. Comprehensive and reliable international migration information (i.e. mostly tabulated statistic data) continue to be poor despite an increasing global interest in migration patterns [3]. The correct interpretation of raw migration data is difficult. Proposed analytical migration models (i.e. gravity, radiation, ...) have demonstrated not to be complete to analyze the complexity of the phenomenon. Moreover, large amounts of prior ground truth mobility data are not available [4]. The combined use of Knowledge Discovery in Databases (KDD), Pattern Recognition and Data Visualization for such purpose can help to analyze and understand better human mobility, and also to discover collective migration patterns.

Thus, the adoption of an adequate e-government platform and procedures for handling migration data, are central to the prediction and analysis of human migration patterns, and consequently for the adoption of proper public policies. Visualization of migration flows is critical for understanding movement of people. Visual supports are needed to acquire knowledge that cannot be directly extracted from raw data. Therefore, effective representations, analysis and visualization of existing migration data remains a challenge to the research community.

The paper is organizes as follows. Section III outlines the architecture of proposed solution. Section IV summarizes the KDD stages applied to build the database of indicators per countries, and also the Pattern Recognition methods that where used for country classification. Section V describes the proposed migration-related data visualization methods provided by the implemented tool. Last Section VI outlines the conclusions of this work.

II. Related work

Recent works have dealt with modeling social phenomena from a variety of techniques, such as complex networks/systems [5], [6], machine learning [7]–[9], and multi-agent systems [10]. Previous works used to determine migration flows the data coming from international development agencies (e.g. United Nations Population Division, OECD, IMF and World Bank, among others) [11], [12]. More recently, there exist also some local studies of human mobility which use information collected from GPS signals of mobile phones [13]. To estimate a function which map indicators from pairs of countries and determine the number of migrants between them, Robinson and Dilkina [4] have used two machine learning based models: gradient boosting algorithms and artificial neural networks, respectively. Other authors have have applied Complex Networks theory to create weighted-directed graphs (using country indicators), which are analyzed and related to migration movements [14]. Although there are not many published works on visualization methods applied to migrations, these tools aim to detect spatial patterns of human movements effectively through space and time [15]. Our work aims to develop a country classification and tool which receives the input data from a normalized database, collected mostly from [16], which includes multiple national socio-economical indicators from several sources, and produce different types of visual representations (i.e. histograms and maps, mainly) which aim to simplify the understanding of migration flows between countries (both in one year and along periods of time). In our approach, previous to visualization stage, several KDD and pattern recognition tasks are carried out for a better interpretation of the information to be visualized. This tool not only contributes to the study of human migration at different geographic and temporal scales, but also provides a starting point for building a more comprehensive analytical platform to examine causes and effects of migrations.

III. Overview of proposed tool

Fig. 1 shows a high-level UML diagram with the involved successive processes (white boxes), inputs, intermediate results and outputs (gray boxes) in the proposed classification/visualization tool. First, the database of countries with multiple socio-economic indicators is manually created from current tabulated data of several sources. All sources used to build the database can be found in the following repository. After that, using Knowledge Discovery in Databases (KDD) methods, the values of country indicators (or attributes) are normalized for the correct application of classification algorithms. This is performed first by setting the countries with a label (i.e. $High$, $Medium$ and Low, respectively) according to its concrete immigration and emigration attribute values. Then, the classification of countries is improved using the remaining attributes (more than a hundred) and using several algorithms from WEKA library [17]. Then, this classification is used to build and visualize indicator-based histograms. The tool also provides multiple types of maps visualizations: distance maps and immigration/emigration maps, which provide insights in the analysis of migration phenomena. In addition to WEKA for machine learning, we have also used the $Scikit-learn$ library. For the code development and creation of graphical interfaces of the tool, several programming languages like Java, JavaScript and Python were used.

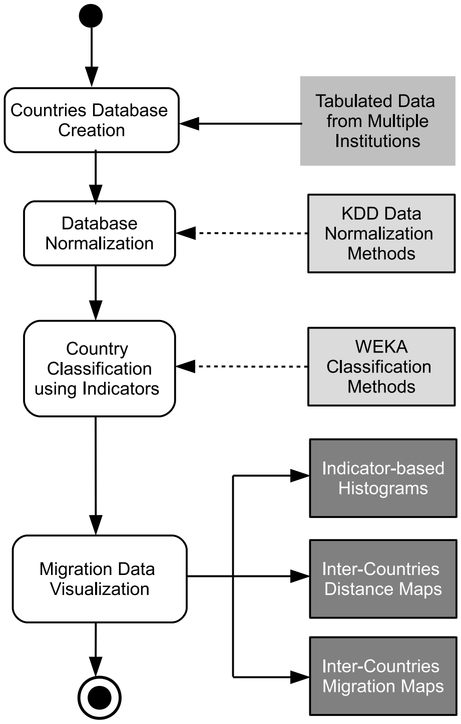

Countries Database Creation
Tabulated Data from Multiple Institutions
Database Normalization
KDD Data Normalization Methods
Country Classification using Indicators
WEKA Classification Methods
Indicator-based Histograms
Migration Data Visualization
Inter-Countries Distance Maps
Inter-Countries Migration Maps

Fig. 1: UML Activity Diagram of Involved Tasks in Proposed Migration Analysis System.

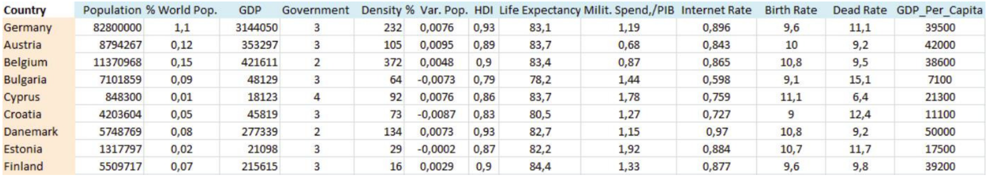
IV. Country classification using socio-economic indicators

This section describes both involved data normalization processing of the database and the classification of countries based on their socio-economic attributes. These two tasks are necessary pre-processing for the different visualization methods included in our tool.

A. Database creation and normalization

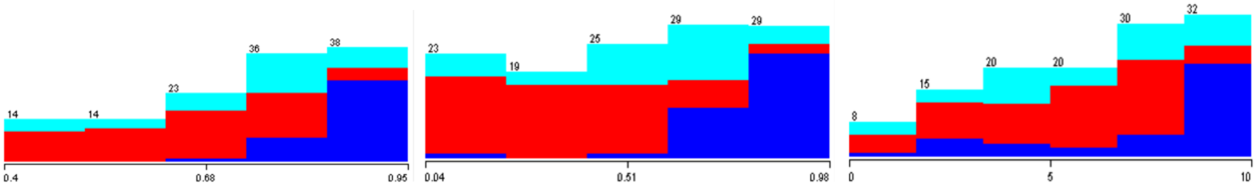
Fig. 2 presents a partial tabular view of our raw database which contains a total of 125 countries (rows) and 118 attributes (columns) per country. These attributes are then normalized to values between zero and one. This dataset has been created manually by combining tabulated information from multiple sources (e.g. United Nations, The World Bank, The World Trade Organization, etc.).

After that, all the features in the database (i.e. Excel table) were normalized to map their values in the interval [0,1]. These normalized data are the converted to .csv files. For such purpose, two normalization methods were used (each ones for different subsets of attributes): $min - max$ and $z - score$,



Country	Population	% World Pop.	GDP	Government	Density	% Var. Pop.	HDI	Life Expectancy	Milit. Spend,/PIB	Internet Rate	Birth Rate	Dead Rate	GDP_Per_Capita
Germany	82800000	1,1	3144050	3	232	0,0076	0,93	83,1	1,19	0,896	9,6	11,1	39500
Austria	8794267	0,12	353297	3	105	0,0095	0,89	83,7	0,68	0,843	10	9,2	42000
Belgium	11370968	0,15	421611	2	372	0,0048	0,9	83,4	0,87	0,865	10,8	9,5	38600
Bulgaria	7101859	0,09	48129	3	64	-0,0073	0,79	78,2	1,44	0,598	9,1	15,1	7100
Cyprus	848300	0,01	18123	4	92	0,0076	0,86	83,7	1,78	0,759	11,1	6,4	21300
Croatia	4203604	0,05	45819	3	73	-0,0087	0,83	80,5	1,27	0,727	9	12,4	11100
Danemark	5748769	0,08	277339	2	134	0,0073	0,93	82,7	1,15	0,97	10,8	9,2	50000
Estonia	1317797	0,02	21098	3	29	-0,0002	0,87	82,2	1,92	0,884	10,7	11,7	17500
Finland	5509717	0,07	215615	3	16	0,0029	0,9	84,4	1,33	0,877	9,6	9,8	39200

Fig. 2: Non-normalized database with indicators per countries.



(a) HDI (b) Internet Availability (c) Civil Rights

Fig. 3: Examples of histograms relating immigration with specific attributes.

respectively. With the first approach, the data is scaled to a fixed range (usually 0 to 1). Using the $z-score$ normalization, the features are rescaled so that they have the properties of a normal distribution with mean equal to zero and standard deviation equal to one. The new respective normalized values of attributes are calculated (using the $preprocesing.normalize$ method of $scikit-learn$ library) as follows:

$$x_{norm1} = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad x_{norm2} = \frac{x - \mu}{\sigma} \quad (1)$$

where: x_{norm1} and x_{norm2} respectively correspond to the $min-max$ and $z-score$ normalized results, x_{min} and x_{max} correspond to minimal and maximal values of the given attribute, and μ and σ are the respective average and standard deviation corresponding to values of an attribute.

These two types of data normalization are not only important if we are comparing measurements that can have different units but also are required for the application machine learning algorithms. In the case of $min-max$ normalization, smaller standard deviations are achieved which can suppress the effect of outliers. This normalization is typically used by neural networks that require data on a 0-1 scale (e.g. multilayer Perceptron). The $z-score$ normalization is used for classification algorithms that require the properties of a standard normal distribution (e.g. logistic regression).

B. Machine learning methods applied to country classification using WEKA

We have first labeled the whole 125 involved countries in our study into the three mentioned categories: $High$, $Medium$ and Low, respectively. The assignment of a given country to one of these classes has been done using the corresponding values of two attributes: $PercentageofImmigrants$ and $Percentageofemigrants$, respectively, where both percentages are calculated with respect to the total population of each country. We analyze the database and set the following thresholds: above 10% for label $High$, between 2% and 10% for $Medium$, and below 2% for Low immigration and emigration, respectively. After that, in order to assign one unique label to each country, we used the rules presented in Table I.

TABLE I: Country label from Immigration and Emigration labels

Immigration\Emigration	Low	Medium	High
Low	Low	Low	Medium
Medium	Low	Medium	High
High	Medium	High	High

After labeling all the countries in this way, we have tested several classification algorithms contained in the WEKA Machine Learning Software Library [17] in order to improve the preliminary classification of countries. For such purpose, we performed a supervised classification using all 116 attributes per country, except the "Percentage of Immigrants" and "Percentage of emigrants" (used for preliminary classification). In that stage, two thirds of patterns (i.e. countries) were used for training the classifiers and the remaining third for testing purposes. The following WEKA classification methods were tested and compared: Naive Bayes ($NaiveBayes$ algorithm), Multilayer Perceptron ($MultilayerPerceptron$ algorithm with parameters: Learning rate =0.3, momentum=0.2, epochs=500), k-nearest neighbors (IBk algorithm with number of neighbors=13), Multinomial Logistic Regression ($Logistic$ algorithm with ridge=0.5) and C4.5 decision tree ($J48$ algorithm with minimum number of instances=2), respectively. The best performance was achieved using the Multilayer Perceptron algorithm, which produced only a 5.88% of error when classifying the set of test patterns.

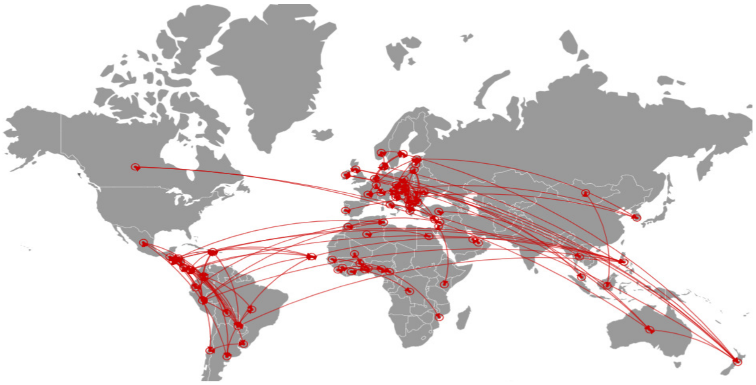


Fig. 4: Distance map using cutoff value of 2%.

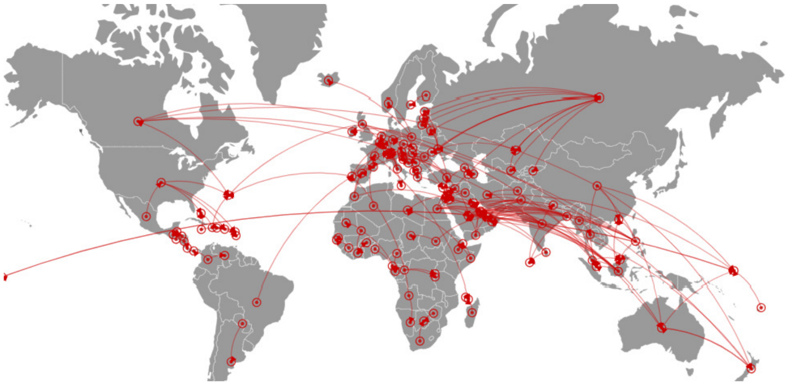


Fig. 5: Immigrants rate map for cutoff values of 1.3%.

V. Visualization of migration data

In this section, we describe the implemented visualization methods for processed attributes from countries in the database, with respect to migrations. Three main visualization models have been considered: global histograms, distance maps and migration maps, respectively. These are described and illustrated in the next subsections.

A. Global histograms relating immigration with specific attributes

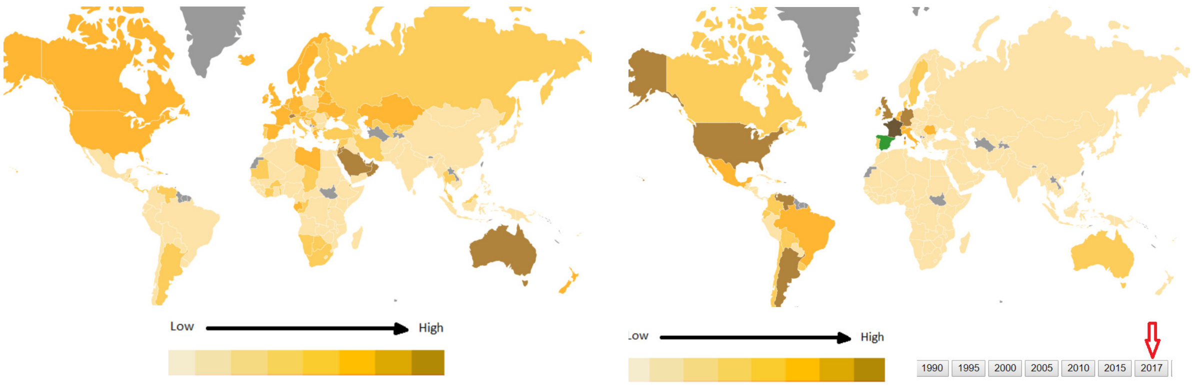
These histograms are computed for each specific country indicator (e.g. Human Development Index or HDI). The values of the attribute for all countries are partitioned in a number of bins, then these are accumulated and related with the migration levels (i.e. emigration or emigration): Low, $Medium$, $High$, respectively. For a better visualization of the histograms, we only considered 5 or 6 bins per attribute. The number countries with Low, $Medium$ and $High$ migration labels are accumulated in each bin, and then the total number of countries per bin are displayed. This way, for each attribute and associated histogram, it is possible to generate a conclusion on the relation of migration with this specific attribute. With this type of visualization it is possible to determine which attributes are most influential on the migration processes. Fig. 3 presents three of these global histograms which relate immigration to respective attributes of HDI, Internet Availability and Civil Rights. For example, with respect to HDI histogram (left subfigure), one can conclude that "most of immigration movements happen in countries with a high value of the HDI attribute" (i.e. countries with high living standards).

B. Distance maps between countries

Another type of representation consists in computing several types of distance maps between pairs of countries. The distance between pair of countries is computed using the normalized attribute values of each country and a distance metric (e.g. Euclidean, Mahalanobis or Chebyshev distance), and considering or not considering the migration features when calculating the distances. This result corresponds to the weight of the connection between two countries. After that, we visualize on a world map only those connections above a cutoff or threshold value. Fig. 4 shows an example of a distance map without migration attributes (immigration and emigration) using Euclidean distance and a cutoff of 2%.

C. Migration maps among countries

Similarly to previously explained inter-countries distance maps, one can create maps relating the number of immigrants with respect to the total immigration in the destination country. Then, using cutoff thresholds (corresponding to percentage values), the strongest connections above the cutoff value are kept. Another interesting type visualization map corresponds



(a) Immigration map

(b) Temporal emigration map from Spain

Fig. 6: Different visualization of migrations provided by the tool

to the number of immigrants with respect to the total population of the host country. Fig. 5 illustrates this second type of map using two different cutoff values 5% and 1.3%, respectively.

It is observed that when using the higher threshold the number of links between countries is reduced. Moreover, nearly 90% of migrations have destination to countries with higher GDP and with higher percentage of Internet users. For the case of 1.3% cutoff value, many new connections appear. These connections are mostly local(i.e about 73% of them) and have destination to destination countries with better conditions with respect to the origin. For example, in the case of Spain, to immigration links appear (with Morocco and Romania) since immigrants from Morocco (in 2017) represented a 1.5% of Spanish population and Romanian immigrants represented a 1.4%, respectively. Another way of viewing the migration figures is to visualize each country with a different color (e.g. where a brighter color means a lower migration and darker one means a higher migration) according to its total number of immigrants or emigrants in the country. It is also interesting to graphically show the evolution of migrations during the successive years for a give country. For such purpose, our tool also includes different visualizations of simple migration maps and temporal migration maps. Fig. 6 illustrates both a simple immigration map and also a temporal emigration map referred to a country (e.g. Spain). The first map is useful to illustrate that many (underdeveloped) countries from Africa, South-East Asia and Latin America are receiving less number immigrants compared to other European and North American ones. The second map shows that most Spanish emigration is directed towards other European countries like France, United Kingdom and Germany. Our tool also allows to visualize the evolution of the immigration for any target country and given a sequence of years.

VI. Conclusion

This paper presented a classification and visualization tool to get insight and to discover non-direct relationships among multiple socio-economic indicators from countries and migration movements. After creating the countries' database, the attribute values of nations are normalized, and then these nations are classified into several categories as a previous stage to visualization of histograms and immigration/emigration maps in different ways. Each visualization type provides a perspective to better understanding the migration phenomenon. In our study, the multilayer Perceptron produced for our dataset the smaller classification error(less than 6%) using several compared machine learning algorithms from WEKA. Using the visualization methods provided by our tool, it is possible to detect that migrations are mostly within-continental, that GDP per capita and percentage of Internet users are important attributes for attraction of immigrants. Moreover, the inter-countries distance is inversely proportional to immigration percentage between countries. As future work, we are interested in improving the pre-processing of our database to reduce the effect of spurious data. Another improvement of the tool will consist in incorporating to it some type of modeling to predict future migration flows.

Acknowledgments

This work has been supported by UAM-Santander CEAL-AL/2017-08, and UDLA-SIS.MG.17.02. The authors gratefully acknowledge the support offered by the CYTED Network: "Ibero-American Thematic Network on ICT Applications for Smart Cities" (Ref: 518RT0559). This work was funded by Spanish project of Ministerio de Economía y Competitividad/FEDER TIN2017-84452-R (http://www.mineco.gob.es/).

References

[1] G. J. Abel and N. Sander, "Quantifying global international migration flows," *Science*, vol. 343, no. 6178, pp. 1520–1522, 2014.

[2] A. A. Tarasyev, G. A. Agarkov, and S. I. Hosseini, "Machine learning in labor migration prediction," in *AIP Conference Proceedings*, vol. 1978, no. 1. AIP Publishing, 2018, p. 440004.

[3] C. R. Parsons, R. Skeldon, T. L. Walmsley, and L. A. Winters, *Quantifying international migration: a database of bilateral migrant stocks*. The World Bank, 2007.

[4] C. Robinson and B. Dilkina, “A machine learning approach to modeling human migration,” in *Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies*. ACM, 2018, p. 30.

[5] M. González, M. del Mar Alonso-Almeida, and D. Dominguez, “Mapping global sustainability report scoring: a detailed analysis of Europe and Asia,” *Quality & Quantity*, pp. 1–15, 2017.

[6] D. Dominguez, O. Pantoja, and M. González, “Mapping the Global Offshoring Network Through the Panama Papers,” in *Proceedings of the International Conference on Information Technology & Systems (ICITS 2018)*, Á. Rocha and T. Guarda, Eds. Cham: Springer International Publishing, 2018, pp. 407–416.

[7] M. González, M. del Mar Alonso-Almeida, C. Avila, and D. Dominguez, “Modeling sustainability report scoring sequences using an attractor network,” *Neurocomputing*, vol. 168, pp. 1181–1187, 2015.

[8] M. González, D. Dominguez, O. Pantoja, C. Guerrero, and F. B. Rodríguez, “Modeling sustainability reporting with ternary attractor neural networks,” in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2018, pp. 259–267.

[9] M. González, D. Dominguez, G. Jerez, and O. Pantoja, “Periodically diluted begnn model of corruption perception,” in *International Conference on Mining Intelligence and Knowledge Exploration*. Springer, 2018, pp. 289–298.

[10] Á. Barbero, M. S. González-Rodríguez, J. de Lara, and M. Alfonseca, “Multi-agent simulation of an educational collaborative web system,” in *European Simulation and Modelling Conference*, 2007.

[11] N. Xiao and Y. Chun, “Visualizing migration flows using kriskograms,” *Cartography and Geographic Information Science*, vol. 36, no. 2, pp. 183–191, 2009.

[12] T. S. Lautenschutz, A.K., “Report on migration and mobility data visualization workshop,” *Technical Report. Zurich Open Repository and Archive. University of Zurich*, 2012.

[13] J. R. Palmer, T. J. Espenshade, F. Bartumeus, C. Y. Chung, N. E. Ozgencil, and K. Li, “New approaches to human mobility: Using mobile phones for demographic research,” *Demography*, vol. 50, no. 3, pp. 1105–1128, 2013.

[14] G. Fagiolo and M. Mastrorillo, “International migration network: Topology and modeling,” *Physical Review E*, vol. 88, no. 1, p. 012812, 2013.

[15] G. Zambotti, W. Guan, and J. D. Gest, “Visualizing human migration through space and time,” 2015.

[16] T. W. Bank. (2019) World development indicators. [Online]. Available: https://datacatalog.worldbank.org/dataset/world-development-indicators

[17] I. H. Witten, E. Frank, L. E. Trigg, M. A. Hall, G. Holmes, and S. J. Cunningham, “Weka: Practical machine learning tools and techniques with java implementations,” 1999.