USING VISUAL ANALYTICS AND INTERPRETABILITY STRATEGIES TO UNDERSTAND THE IMPACT OF INPUT VARIABLES ON INDEXES DERIVED FROM MUNICIPALITY (URBAN) DATA SETS

by

Balaji Dhakshinamoorthy

Submitted in partial fulfillment of the requirements
for the degree of Master of Computer Science

at

Dalhousie University
Halifax, Nova Scotia
April 2020

The thesis is dedicated to my family who have been the great support to me for this long journey.

Table of Contents

List of Tables

List of Figures

Abstract

Composite indices have been widely used in several domains as a measure to describe abstract concepts through the combination of variables. The current approaches for index creation and analysis do not have a comprehensive visual interface to enable the use of external information to support interpretation. We propose a visual analytics framework that places users in the loop for creating and interpreting indexes. It helps users to compose an index with the flexibility of determining a weight for the linear combination of indicators. For the interpretation, we use regression analysis to provide explanations for indexes from both internal and external variables. we demonstrated use-case scenarios using crime and demographic datasets to show the benefits of our interface for decision-making tasks at the municipal level. we validate our results through a comprehensive user evaluation, showing that most users reach similar conclusions when using our framework to execute analytical tasks.

List of Abbreviations Used

AutoML Automated Machine Learning

HRMD Halifax Regional Municipality Demography

HRPC Halifax Regional Police Crime

LIME Local Interpretable Model-Agnostic Explanations

PCA Principal Component Analysis

Radviz Radial Coordinate Visualization

Acknowledgements

I am very thankful for the guidance from my professor Fernando Paulovich. He has been the great pillar of support for my complete Thesis Journey. It was a great opportunity to work under his supervision. I have to mention two other people from the Visual Analytics and Visualization lab whom I owe my sincere gratitude for their contribution as a mentor for shaping the crucial components of this thesis. I would take this opportunity to thank Martha Dais Ferreira and Leonardo Christino for their critical suggestions that helped me improve the quality of the thesis. I would also like to thank my family for their constant encouragement and moral support that has helped me throughout the journey of this thesis.

Chapter 1

Introduction

Composite indexes have been widely used in various domains to compose the sets of variables or indicators that reflects concrete information in to a single measure for the purpose of summarizing the abstract concepts [42]. For instance, the well-known Human Development Index combines life expectancy, education, per capita income, and other factors [6, 44] to summarize and rank human development of the countries. In most of the cases, the indicators used to specify an index are multidimensional and complex. They can contain different types of data, like categorical, numerical and ordinal [42, 32, 4]. The index composition is usually not a simple formulation and it involves different procedures to compose indicators.

The most common technique applied to compose an index is the Principal Component Analysis (PCA) [24]. PCA projects the multidimensional variables into latent factors, called principal components, that can explain the maximum variance of the data [20, 6, 41, 26]. Despite its popular application, PCA is not suitable to compose indexes when the indicators are uncorrelated [44]. Further, the PCA does not support user intervention, i.e., user defined weight for individual indicators is not captured properly by PCA for index composition. Therefore, the indicator with maximum weight and lower data variance is neglected in the process of index composition. Further, the technique that applies PCA for index composition uses the weight obtained from the eigen values of the principal components to compose the index and does not consider the weight defined by the user [44, 6].

One recent aspect of a composite index field that has gathered attention to support decision-making is interpretability. Interpretability is the analytical task of comprehending an index from the input indicators or external variables, also called explainers [3, 38, 44]. Currently, the most common data mining strategy for interpretation is clustering analysis. The indicators are grouped as clusters based on their similarities, and those cluster groups are used to interpret the groups defined by the interval of

index. Further, they also identify the indicators from the factor loadings with maximum coefficients and use them as input for clustering to interpret the index [38, 44]. Visualization techniques have also been applied to support interpretation, helping the exploration and understanding of indexes in multiple geographical regions [3]. Although representing an evolution, these strategies only support the analysis of an index from the indicators used in its composition. It is not possible to execute tasks that involves external information. For instance, to interpret the violence index using demographics information. [40] presented an approach for index interpretation from external variables using regression analysis. Although an intriguing study, it is limited for the notion that it needs expert opinion to identify the appropriate external indicator to interpret the index. Our system addresses this limitation with the interactive visual interface for the selection of any external variables to interpret the index without the need for Machine Learning Expert. We used the AutoML approach [7] to approximate the relation between explainers and index and LIME [36] to identify the appropriate explainers.

In this dissertation, we propose a visual analytics framework to address the limitations, placing users in the loop for creating and interpreting indexes. For the index creation, we propose an interactive visual interface to compose an index from a linear combination of indicators, with user defined weight for each indicator. For the index comparison, we propose an interactive interface to help users to visually compare two indexes with the support of piechart and geospatial choropleth visualization. For the index interpretation, we use regression analysis supported by a machine learning interpretability approach and visualization techniques to provide explanations for indexes from the variables used in their composition and multiple external variables. We use an Automated Machine Learning (AutoML) [7] approach to support those tasks, so that there is no need for any machine learning expert for the process.

Hence, the Visual Interface will be effective in the composition of the index and produces efficient model interpretation for the explainers. The interactive user interface reduces the burden of choosing the right explainers for identifying the reason behind the generated index. The intuitive nature of the user interface makes the tools accessible for end users and help them to understand their generated index without the aid of the expert.This makes the tool powerful and with some instruction and

exposure to the interface, end users will find it comfortable to navigate through all features of the interface.

In summary, the main contributions of this dissertation are:

- A strategy for index composition and analysis that takes into consideration user knowledge;
- An approach based on AutoML and regression models to explain composite indexes given a set of external indicators;
- A visual analytics framework to support users on designing, exploring, and understanding indexes interactively.

Chapter 2

Related Work

An index is a composite measure to evaluate and understand multidimensional information, aggregating multiple indicators to summarize different perspectives of complex data sets. It helps to identify the strength and weaknesses of a system under analysis, comparing large collections of data instances through a ranking strategy. In general, indexes are deemed as the most relevant measure for comparing multivariate data, especially when there is no standard tool for evaluation [42]. Their application includes measurement of progress in human development, human poverty, and various social and economic indicators [6, 44]. The index is also the most common tool for communication among policymakers and municipalities [32, 4].

2.1 Index Creation

Principal Component Analysis (PCA) [24] is the most common technique to create an index. PCA projects the multidimensional variables into latent factors, called principal components, that can explain the maximum variance of the data. Principal components are obtained from the multiplication of the original indicators with the corresponding eigenvectors generated from the covariance matrix of indicators. The resultant principal components with eigenvalues greater than 1% are then linearly combined with their corresponding eigen values as weights to produce the index [20, 6]. [41] used PCA through an alternate approach to compose an index. They followed a similar procedure by identifying the principal components with eigenvalues greater than 1% and linearly combined them using predefined loading weights for each indicator to compose an index. [26] followed a similar approach, by reducing he number of components in the factor loadings from the variance information captured by principal components as a threshold measure, ignoring the components with associated eigenvalues that are lower than 1.

Although a commonly applied technique for index composition, PCA is not appropriate when the input indicators are uncorrelated [44, 41, 21, 33]. PCA is a linear transformation technique that transforms the group of high-dimensional variables with certain correlation into lower-dimensional uncorrelated components. Hence, the analysis requires a certain correlation among input indicators, of at least 0.3 [21]. Therefore, scenarios that present uncorrelated indicators demand a need for an alternative approach for index composition. In our case, as the user selects the indicators, it is impossible to obtain the recommended correlation among all combinations of indicators. Also, as PCA defines the importance of indicators according to the data variance, the user is not able to include their knowledge in the index creation, which may lead to results that do not match user expectations. Henceforth, in our framework, we allow the users to select indicators and define their importance for the index combination, with the flexibility to adjust the importance through visual interaction and generate indexes for any combination of indicators without any impact from correlation among them. We address this issue through a simple weighted linear combination of indicators.

User-controlled weighted linear combinations have been used in literature [45, 18]. [45] used this strategy to identify agriculture vulnerability for climate change in nordic countries. The generated index is visualized using a geospatial visualization as followed in many index-based composition techniques for geographic areas [45, 18, 5].Once the index is composed, usually, there is a necessity for a visualization system to understand the composition of indicators for each geographic area. [18] formulated this principle of visualization of index composition to construct an interactive integrated index. The user can visualize the final index based on their interaction on indicator composition. The indicator contribution for the index is represented using parallel coordinates [23]. They have also used the Sankey diagram [37] to present the composition of an indicator for the generated index. They designed the system with the capability to generate or visualize indexes based on user interaction. This procedure of understanding index based on the user interaction with the geographic area is similar to our approach of understanding index composition. Their system lacked the functionality to compare the indexes generated by the users and the lacked the ability to interpret the index from external sources. Our system addressed those

limitations, with some advanced work having functionality implemented for index comparison and index interpretation using independent external variables that are not used in its composition.

2.2 Index Interpretation

A challenge often faced by researchers is the composite index interpretation, which attempts to identify the appropriate measure to understand the generated index scores [38, 44, 3]. Such interpretation supports the understanding of why a specific instance (or geographical region) have lower/higher values given a set of indicators (or attributes). In this context, many researches interpret an composite index using clustering analysis applied to the original indicators [38, 5]. In this scenario, [38] used a hierarchical clustering on the input indicators to cluster them into 4 different groups of municipalities based on their well being. They found that municipalities grouped using hierarchical clustering are identical to the municipality groups defined by the composite index generated using the PCA technique. This interpretation helped them to trust the generated well-being index for the municipalities.

Another study using hierarchical clustering was proposed by [44], which aims to identify the groups of municipalities based on socio-economic indicators. They identified the crucial variables that make a maximum impact on the major principal components for the index composition. After the identification of those indicators, they applied the clustering to group the municipalities to interpret the index. The index generated for the set of indicators could be interpreted using independent variables that are not used in the composition. This analysis requires the identification of external variables that could explain the indexes generated for the geographic areas.

In some studies, index interpretation was conducted using regression analysis, for instance, [40] presented a study for index interpretation using external variables. The idea was to interpret the high-density crime areas of Halifax through the approximation relation from demographics variables of the same city. They adopted the least-squares regression method to analyze different demographics variables for reasoning about high and low crime density areas. They conducted multiple regression analysis for each demographic variable and determined the variables that explained

the maximum variance of the crime density areas. They were able to identify the demographics characteristics, like income or education, that contribute to higher crime rates in different neighborhoods. Although an intriguing study, they are limited to the view that they need an expert opinion to identify the suitable external variables that can explain the approximation relation with crime density areas.

In order to address this issue, we advanced that limitation by generating the approximation model for any combination of external variables with the generated index without the need for expert intervention. We adopted an AutoML strategy [17] to identify the best approximation model for the selection of external variables with the index. AutoML automates the process of identifying the best approximation model for selected independent external variables and composed index through the pipeline combination of different machine learning algorithms like Support Vector Machine [31], Decision tree [39], and others. On the identification of the right model for the scenario, we employ the LIME [36] approach to figure out the contribution of external variables for the generated model. That contribution information is presented visually on top of the corresponding geographic area based on user interaction. This interpretation strategy can help users to interpret the index from the external variable perspective.This technique makes the interface more advanced than other techniques proposed for index interpretation using clustering analysis [38, 44, 3].

2.3 Index Comparison

Visualization techniques have been applied to understand the impact of different indicators for an index and to compare indexes generated using same indicators for different years. [3] compared different visualization techniques for understanding the index composition of six different cities generated for a specific period of years. They found that Radial Coordinate Visualization (Radviz) [3] requires more time and effort to interpret the indicator composition for those composite indexes than the Circle Chart [8] and Flower Chart [8]. They concluded it based on the study conducted among users for the identification of the best visualization model for index composition. However, the approaches mentioned above have some issues in terms of visual scalability, especially with the addition of more geographic areas. It makes them limited to compare and display only a few instances at a single time.

We address that limitation by limiting the index comparison to two indexes at a single time. The selected index differences are displayed on the geospatial visual interface, that is flexible enough to represent any number of geographic areas. Then, the pie chart will be displayed to visualize the indicator composition for the selected indexes. This specific technique using pie chart to visualize the indexes was successfully adopted before to visualize the indicator composition of different well-being groups [18]. This visualization makes the understanding of the composition of different indexes easier, as indicated by [18, 19]. Hence, our framework with this combined approach of a proper index composition technique and various interpretation strategies will make the decision-making process more convenient.

Chapter 3

Methodology

In many scenarios, the evaluation and understanding of multidimensional data are performed using indexes, which are composite measures that summarize different perspectives of variables or indicators [42, 20, 6, 41]. In this context, the proposed methodology aims to support the creation of indexes and the interpretation of the relation of a given index and external indicators, that is, indicators not used to compose the index. The goal is to assist the understanding of why specific instances (or geographical regions) have lower/higher indexes from a set of indicators [38]. Many researchers explored clustering analysis to conduct this interpretation [38, 44]. Our approach is different from them for the reason that our goal is to understand how external information can explain an index.

To implement the interpretation, we use regression analysis to model the relation between the given index and a set of user defined external indicators [34, 22, 29, 43, 13, 12, 11], followed by interpretation strategies to understand the model. As most users of our platform are non-experts in Machine Learning, an Automated Machine Learning (AutoML) framework is applied to compute the appropriate model and its optimal parameters without the need for expert or user intervention [17, 16, 7, 35, 14]. In this context, we use the Tree-based Pipeline Optimization Tool (TPOT) [35] to generate the regression models. In sequence, to support the analysis of the produced model, we use the Local Interpretable Model-agnostic Explanations (LIME) [36] as the interpretation tool.

Figure 3.1 presents an overview of our approach. Initially, an index composed of various indicators is created, in which a user selects indicators and weights to specify their importance (1). The user can then analyze the created index, exploring how each indicator contributes to the index value assigned for a specific geographical area (2). The user can compare a pair of indexes to explore the differences between two different compositions (indicators + weights) (3). Then, the TPOT tool is applied

to approximate a model that represents the relation between the produced index and external indicators selected by the user, in this dissertation called explainers (4). In this process, the index is the dependent variable, and the explainers are the independent variables used to create a regression model that approximates (or predicts), as much as possible, the given index. Finally, LIME is applied to explain the relation of the index value and explainers for a specific geographic region(5). In the following sections, we detail each of these components.

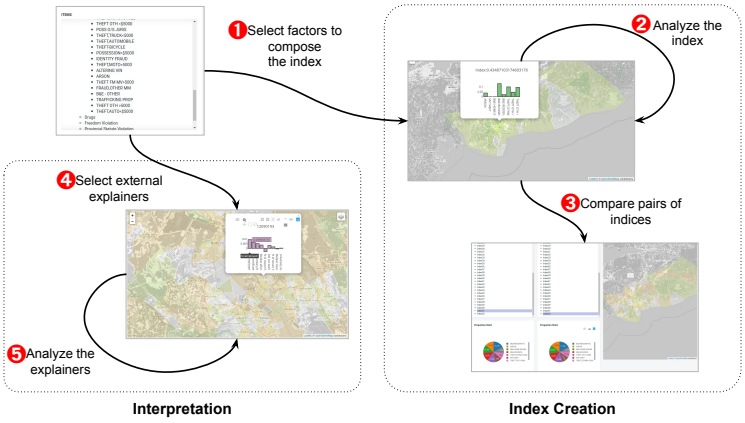


Figure 3.1: Overview of our approach to create and explore composite indexes. The user starts combining indicators to create an index. Then,our system applies regression analysis to model the relationship between the index and external indicators or explainers. Such relationships can then be evaluated using a model agnostic strategy for model interpretation.

3.1 Index Creation

In our system, the indicators used to generate an index are selected by the user, who can evaluate different combinations to check the contributions of each indicator for the index value of a geographical area. The user assigns the weight associated with each selected indicator, indicating their importance for the final index.

Principal Component Analysis (PCA) [24] is the commonly applied technique to produce an index [20, 6, 41]. However, after some experiments, we observed that the principal components' data variance might not explain the user expectations, leading to indexes that do not match users' perspective and knowledge. Further, according to [21, 44], PCA is not an appropriate technique for uncorrelated indicators, i.e., when

the correlation among indicators is less than 0.3, proving that PCA is not suitable for all inputs. These studies led us to devise a simple linear combination strategy to produce an index, since this allows users to define the importance of each indicator and can be used in any domain under analysis.

In our approach, the data set is a frequency matrix, in which the rows denote geographic areas and columns denote the indicators selected by the user for the geographic area. Based on the selection of indicators, the frequency of the indicator occurrence for each geographic area is computed and updated as the input data matrix. Equation 3.1 defines the linear combination used to compose an index, in which I_i is the index for the geographical area i, $x_{i,j}$ represents the instance i of indicator j normalized by the maximum value of the indicator $\max(x_j)$, w_j is the weight associated for indicator j, and k represents the size of selected indicators. The threshold of the weight is determined from two different procedures: i) if the input data source has information related to the importance of indicators, it can be imported as a weight threshold after applying normalization; ii) otherwise a uniform normalized weight threshold is applied for all indicators.

$$I_i = \sum_{j=1}^{k} \frac{x_{i,j} w_j}{\max(x_j) \sum_{j=1}^{k} (w_j)} \tag{3.1}$$

Equation 3.1 produces the composite index I_i for all the geographic areas i, where I_i is in the range $[-1, 1]$. The user can determine the direction of weight before the index composition, which can be either positive or negative to signify the impact of the specific indicator for the final index score. It is worth mentioning that this strategy allows the user to select negative weights to select an indicator that impacts negatively for the final index.

3.1.1 Index Visualization

To visualize the generated index we use the well known choropleth metaphor, representing the geographic areas in specific patterns or shades based on the index values. For our interface, we use a color pattern from red (negative values) to green (positive values), with white for zero value. Also, we divide our interval $[-1, 1]$ into nine bins, 4 for positive values, 4 for negative values, and 1 for zero. The continuous values of

I_i are then assigned to these bins so that all values in one bin receive the same color. It makes the visual interface more interpretable to diagnose the low and high index areas appropriately.

In addition, a bar chart is used to represent the rank of the geographic areas for the generated index. The user can interact with the rank chart to identify the geographic area that belongs to a specific index value based on mouse hover operation. Then, the appropriate geographic area is highlighted in the map visualization. The user interaction on that highlighted area pops-up a bar chart that explains the contribution of each indicator for the index I_i that belongs to the area i. The size of the bars in the bar chart for the specific geographic area is computed as

$$G_{i,j} = (x_{i,j} * w_j) \tag{3.2}$$

where $G_{i,j}$ is the bar size of the indicator j in the geographic area i, w_j represents the weight of each indicator j, and $x_{i,j}$ represents the frequency of occurrence of the indicator j for the geographic area i.

This gives an additional perspective of how an index I is composed for each geographic area, helping users to break down the generated composite index and identify each indicator contribution. Such visualization also assists the users in the reasoning of why a specific index is higher/lower for that geographic area. Such visualization improves the interpretation of the index, making it more transparent and easier to identify the reason behind the index score for any specific area.

3.1.2 Index Comparison

Another feature for the index analysis is the pairwise index comparison, which makes the interpretation of index composition more convenient. Users can compare two composed indexes I and I' produced by different indicators or/and weights, visualizing the differences between them. Two different visual representations assist this process. First, the pie chart shows how each index is composed, and the second, choropleth map visualizes the difference between the selected indexes, i.e., mapping the difference $I - I'$ to color. For the geographic areas with identical values of indexes I and I', the choropleth pattern is white; otherwise, the shades of colors indicate its difference strength and to signify if it is a positive or negative difference. The patterns of red

indicate the region where Selected Index I has higher values than Index I'. Similarly, patterns of green indicate the geographic regions with higher values for the selected Index I'.

After the choropleth visualization, the user can further explore each geographic area to interpret their composition for the two different indexes. After the user click interaction on a specific geographic area, a small screen is popped up with the overlay bar chart representing an indicator contribution for the composition of two different indexes for that area. The bar chart shows the proportion of all indicators for the two different indexes with different color representations to differentiate them. The indicators that are unique to each index are stacked alongside them without any overlap. This visual representation further reinforces users to understand the index (or indexes), helping identify the proper combination of indicators and to study the spatial distribution of any two indexes with minimum effort.

3.2 Index Interpretation using Explainers

The interpretation of the generated index I from the selection of external indicators, also known as explainers, uses regression analysis to create a model that approximates the relation between the explainers and the index. Regression models calculate the function that uses independent variables to approximate the dependent variable, employing statistical methods [34, 11]. If there is an approximate function that can relate the explainers with the index, they are chosen to explain the index. From the approximate function, the independent variable's contribution in a positive or negative direction for the dependent variable index is used to explain the index [12].

3.2.1 Regression Analysis

To support the regression analysis, we design an adaptive system that accepts any selection of external variables from the user for the explanation of the generated composite index. As the user may not have enough knowledge to select the best model or parametrization for the regression analysis, we include in our framework an Automated Machine Learning (AutoML) tool, an automated process to identify a suitable model for the data. It conducts a function approximation and hyper parameter optimization for different models, evaluating them with cross-validation. Eventually, an

suitable model is produced that can fit the independent variable to the dependent variable.

In our approach, the AutoML technique applied was the Tree-based Pipeline Optimization Tool (TPOT) [35], which uses genetic programming to identify the most suitable model for the data under analysis. Such a technique presents good results for the regression task compared to other AutoML techniques, which motivates its application in our scenario [17]. TPOT evaluates different models for a given data set, with the goal of minimizing the error given by the negative value of the mean squared error (MSE) [35]. Equation 3.3 defines the MSE, in which y is the observed value represented by the index produced, y_{pred} is the prediction provided by the model considering the explainers, and n represents the size of instances associated with the region. Such measurement corresponds to the model residuals, indicating the expected value for the error loss. Thus, we decide to use MSE to measure all models applied for this task [25, 11, 22].

$$MSE = \frac{\sum_i^n ((y_i - y_{pred})^2)}{n} \tag{3.3}$$

The coefficient of determination R^2 is also used to verify the produced model that defines an approximation relation between the independent and dependent variables, indicating how much variance of the index can be explained by the external variables [28]. Equation 3.4 defines such coefficient, in which $\bar{y_i}$ is the average of the observed value y_i represented by the index. It indicates how much the model maintains the variance of the data, providing a good explanation of the data variability when R^2 is closest to 1 [25, 11, 22]. It is worth mentioning that, in some cases, good models can provide lower R^2 values, meaning that the variance of the input data was not kept [28].

$$R^2 = 1 - \frac{\sum_i^n ((y_i - y_{pred})^2)}{\sum_i^n ((y_i - \bar{y_i})^2)} \tag{3.4}$$

3.2.2 Regression Model Interpretation

Identifying the contribution of the external variables for the approximation model is the final stage involved in the interpretation process. The presentation of how explainers can be associated with the indexes of each dissemination area improves

the user understanding of the generated index. We apply the Local Interpretable Model-agnostic Explanations (LIME) [36] method to support this postulation since it is a good approach to calculate the explainers [30], providing secure information and auditing facilities to the users of machine learning models. Through LIME, we can identify the positive and negative impact of each explainer for the generated index.

LIME was proposed to interpret any machine learning model, exploring the local area of a prediction through a surrogate model. LIME conducts a local minimization of the loss function $\ell(f, g, w_x) + \Omega(g)$ to recover an instance in the original representation, in which f represents the model to be interpreted, g corresponds to a surrogate model, $\Omega(g)$ is the complexity of model g, x is the instance under analyzes. In this case, ϵ is used to measure the proximity between x and the instance z provided by the surrogate model, which is the recovery sample used in the original model to obtain the label associated to it. Then, the goal is to minimize this loss function locally to learn the local behavior of the interpretable inputs. In our scenario, x represents the explainers for a specific geographic area, and z corresponds to the explainers reconstruction that should be approximated to the index, which is provided by a simpler surrogate model improving the understanding of the local regression model produced by TPOT.

In general, LIME selects a set of instances around the area of a query sample to generate a new explainable model for that local area, which permits to expose the local behavior of a complex model. Moreover, the loss function is based on the surrogate model's ability to replicate the global model. Then, when the loss function is minimized, it indicates the trust level achieved by the local surrogate model for the global model. LIME presents excellent potential in terms of helping users and developers to explore and understand machine learning models, increasing its reliability. In our scenario, after passing the required information to the LIME, it generates the probability of the contribution of each external variable for the model. Those probability scores are based on the weight of each explainer that are used as parameters in the local surrogate model.

3.2.3 Interpretation Visualization

After calculating of explainers' contribution to the index using LIME, there is a need for an appropriate visualization to represent the information for each geographic area. Like the visualization for interpreting the index composition, the explainer visualization is implemented on top of each geographic area. The map interface has two switchable layers, named "Index Composition" and "Interpretation". For the "Index Composition", the bar chart visualization displays the indicator proportion for the index. For the "interpretation", the bar chart visualization represents the probability score obtained from LIME for the approximation function that relates the explainers to the index. So the user can visualize both the index composition and its explainer composition, switching the two map layers like the metaphor adopted by Google Maps[2]. The user click interaction on the specific area displays the bar chart that represents the indicator or explainer composition as per the layer selection.

Chapter 4

Results

In this section, we present the use case of our tool, and few scenarios that can demonstrate the efficiency of the tool showing how it can support index creation and index interpretation. For the first part, we describe the data sets used in this dissertation.

4.1 Data set

We tested our tool on different data sources for the Halifax Regional Municipality. Of the two data-sets, one was used to create the composite index and another one was used as the source for external indicators to interpret the index. For the composite index indicators, we used the data-set of Halifax Regional Police Crime (HRPC) for the year from 2005 to 2017. For the explainers, we used the open dataset from Statistics Canada which contains the demographic information of the Halifax Regional Municipality for the year 2016. To maintain the uniform geographic area for both the data sets, the data transformation technique is applied on the input variables to update the point wise location information in to the dissemination area defined by the statistics Canada for the year 2016. The point inside a polygon algorithm is applied to transform those point wise location information in to specific geographic boundaries. After the transformation, the total unique dissemination areas for both data-sets stand around 601 [9].

The Halifax Regional Police Crime (HRPC) dataset is gleaned from the Uniform Crime Reporting Incident source [10] organized by the Canadian Center for Justice Statistics [1]. HRPC was created to measure the incidence of crime in Canadian society, containing detailed information on crime incidents, accused persons, and victims. In our work, crime incidents and their occurrence are the variables used in experiments, ignoring the other information about the accused persons and victims. The dataset contains the following attributes: i) date and time of incident; ii) the type of crime committed represented by a 9 series of numbers, for instance, 1000 represents

the crime committed against a person; iii) the severity weight for each crime with a minimum of 5 and a maximum weight of 7554.942; iv) the description of the crime incident reported; and v) the dissemination area of the reported crime location. In this dissertation, we structure the data as a matrix, in which rows represent the DAUID, columns correspond to the different crimes reported, and the values are the total occurrence (frequency) for each crime reported in an associated dissemination area.

The second dataset is gathered from the Statistics Canada website [9] that describes the demographic information of the Halifax Regional Municipality Demography (HRMD). The dataset contains the demographic information of the 601 dissemination area in the Halifax Regional Municipality. They are categorized as Population and Dwelling, Age, HouseHold and dwelling, marital status, family details, languages, income and immigration details, education, labor force status, work activity, commute details, transit mode, and mobility status. Each category has subdivisions detailing the data distribution. Like the HRPC dataset, we structure the HRMD as a matrix, in which the rows represent the DAUID, whose columns correspond to the categories, and the occurrence values that are featured along with the subcategories for each dissemination area. It is worth mentioning that, in the case of the variables with hierarchical structures representing its position, their identity is maintained in the data-set with the merged annotation name describing its hierarchical structure. The necessary transformation is applied to recognize those variables in the data-set after the request from the user.

In summary, the crime data-set is used to create the index. And, the demographic dataset is used to interpret those generated indexes, so that we can identify any possible relation between these two data sets. To evaluate the results produced by the system, we consider the baseline produced by Statistics Canada [1] to understand the crime density of the Halifax Regional Municipality for the year 2001 [40]. In this study, they used regression analysis to verify if there is an approximation between the specific independent variable and crime density value of a geographic area. They used the null hypothesis formulation to relate those independent variables with the crime density values and identified the corresponding independent variable that matches the

[1]Statistics Canada produces statistics about Canada to support a better understanding of the country stats. More details can be found in `https://www.statcan.gc.ca/eng/about/about?MM=as`

probability value, part of the statistical hypothesis test. They recognized the impact of those independent variables from the function parameters of the regression model to explain the index. We will be simulating two scenarios for the evaluation of tools like the scenarios used by the statistics Canada to understand the high crime density areas of the Halifax Regional area for the year 2001[40].

We start our discussion with the use case of our tool, followed by the three scenarios to evaluate the efficiency of the tool in index creation and interpretation. As mentioned earlier, the first two scenarios replicate the experiments conducted by the Statistics Canada to interpret the high crime density areas of Halifax for the year 2001 [40] and the last scenario simulates the tool usage by the expert to understand the drug related crime index of Halifax.

4.2 Index Creation

Some existing strategies for index composition depend on Factor Analysis or Principal Component Analysis (PCA) [41, 44, 21]. PCA [24] is one of the commonly applied technique for Index Composition [41, 44, 20, 6]. Shortly, PCA combines indicators according to their variance, generating indexes that represent distribution aspects of the data. Despite its popularity, in our scenario, PCA is unsuitable because unlike other approaches we determine the weight for our indicators from knowledge of external users and not from the principal component explaining maximum variance. In our scenario, the weights associated with crime indicators, called severity indexes, are defined by experts through the interaction with the system.

Moreover, PCA works best when the collection of indicators has a certain correlation among them before applying the dimensionality reduction operation [21, 44]. In our scenario, the indicator selection is dynamic and determined by the end-user, which does not guarantee a reasonable correlation, being contrary to the most application of index composition where they considered the correlation of input indicators before applying PCA. From this aspect, we decide to calculate the correlation degree among selected indicators and proceed with the PCA if there is some reasonable correlation among them, in which 30% is considered acceptable[44]. Such a concept aligns well with the procedure of the PCA, where the goal is to orthogonally transform the set of possibly correlated indicators into the set of linearly uncorrelated components.

Some experiments support that correlation assumptions, HPRC indicators that are correlated provide the index like linear combination. However, in some cases, the indicators do not present a substantial correlation, confirming that our linear combination is more adequate for the purpose. The addition of user-defined weight to the input indicators could change the correlation factor. Most PCA based Index Composition doesn't consider this user-defined weight, they either used Eigen values as weights or the combination of Eigen values and predefined indicator weights [21, 44].

We attempted different approaches to compose the indexes, including variations of PCA and simple linear combinations. We observe that the PCA technique tends to ignore crime indicators with low frequency but high severity (e.g., murders), which do not represent the user expectancy in most cases. Such behavior occurs, because the occurrences of high severity crimes are small when compared to low severity crimes, tending the PCA components to prioritize low severity crimes.

Table 4.1: correlation obtained for the selected indicators *Murder,1st degree*, *Robbery of firearms*,*Shipping Act*,*Luring a child* and *Sex Aslt-wpn/CBH*. The values that are greater than 0.30 represents the good correlation between indicators.

	Murder 1st degree	**Robbery of firearms**	**Shipping act**	**Luring a child**	**Sex Aslt-wpn/CBH**
Murder 1st Degree	1	−0.041225	−0.08327	−0.10732	−0.11462
Robbery of firearms	−0.04122	1	−0.01328	−0.03837	−0.05242
Shipping Act	−0.08327	−0.01328	1	−0.07752	−0.10589
Luring a child	−0.10732	−0.03837	−0.07752	1	−0.14652
Sex Aslt-wpn/CBH	−0.11462	−0.05242	−0.10589	−0.14652	1

Therefore, a simple linear combination to produce the index is more suitable for this particular scenario. In this experiment, we presented a set of indicators that did not achieve a reasonable correlation. The indicators are *Murder,1st degree*, *Robbery of Firearms*, *Shipping Act*, *Luring a child*, and *Sex Aslt-wpn/CBH* for the year

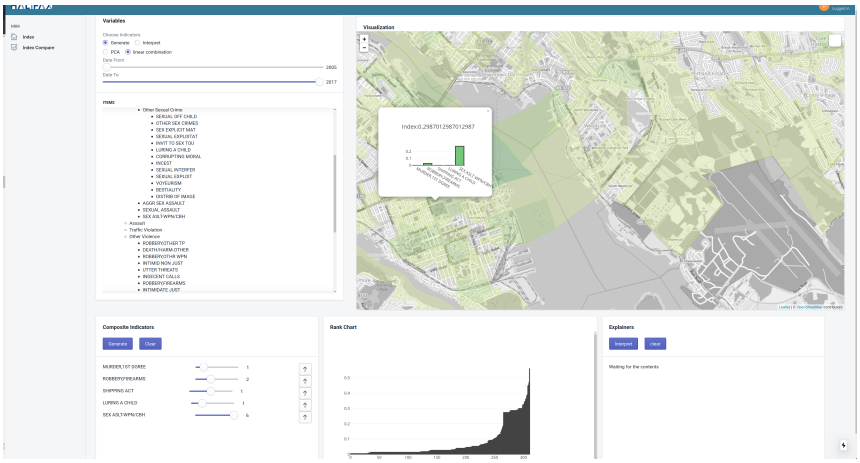


Figure 4.1: Composite index produced by our simple linear combination with the Indicators *Murder, 1st degree*, *Robbery of Firearms*, *Shipping Act*, *Luring a child* and *Sex Aslt-wpn/CBH*. Light to dark green areas indicate the intensity of Index Value. The Bar chart represents the rank of all geographic areas in ascending order.

2008-2017.Their weights are 1,2,1,1 and 6 respectively.Table 4.1 represents the correlation between the selected indicators. We applied the two different methodologies to generate the composite index for the selected indicators.

We found that there is a substantial difference in composite index generated for PCA and our simple linear combination. We identified the area with the higher index value to compare the results produced by our linear combination technique and PCA. Figure 4.1 illustrates the index value (0.298) with *Sex Aslt-wpn/CBH* as the indicator that contributes most for this value. Figure 4.2 represents the crime index generated for the same geographic area produced by PCA. The index value is 0.20, and *Robbery of Firearms* is the indicator with the maximum contribution. It ignores the contribution of the high priority indicator *Sex Aslt-wpn/CBH*. This demonstrates the limitation of PCA when there is not a reasonable correlation between input indicators and its inability to prioritize a given indicator. According to the police severity index, the indicator *Sex Aslt-wpn/CBH* is a severe indicator. In general, PCA might not be the appropriate technique if the input selection contains indicators having correlations of less than 0.3 [21, 44]. This demands us to adopt an approach with more flexibility to handle any indicator combination and expert-defined weights. Therefore, despite its simplicity, the linear combination shows to be an ingenious approach for scenarios where there is a dynamic selection of input indicators and input weights determined

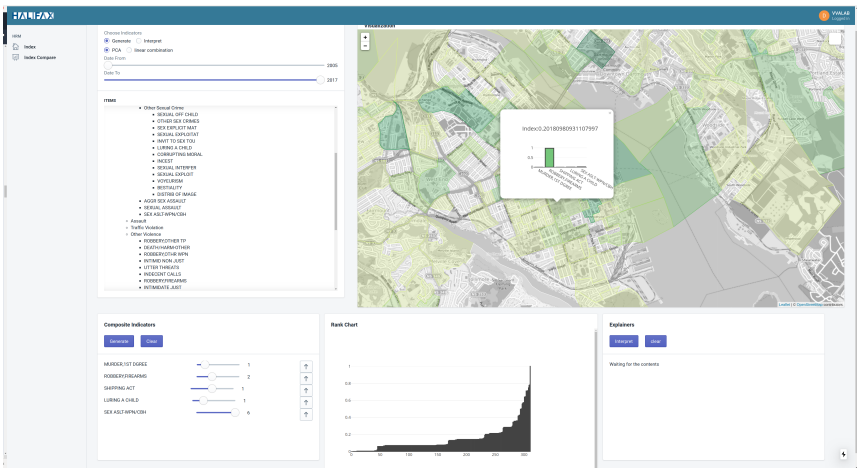


Figure 4.2: Composite index produced by PCA with the Indicators *Murder, 1st degree*, *Robbery of Firearms*, *Shipping Act*, *Luring a child* and *Sex Aslt-wpn/CBH*. Light to dark green areas indicate the intensity of index values. The bar chart represents the rank of all geographic areas in ascending order.

by the expert.

4.3 Index Comparison

Once different indexes are created, they can be analyzed to check the differences and similarities. Figure 4.3 shows an example of such analysis, revealing the differences in the distribution of the previously generated indexes. This visualization can aid in understanding the index composition in two different ways. The first scenario is to identify the index distribution for the geographic area that are generated across the years. In the case of composite index generated for the different years with the same indicator combination, this procedure helps identify the geographic areas with similar distribution and areas with both increasing or decreasing values for that specific index composition. So, it narrows down the identification of areas, which needs more attention for that specific index. The second scenario is for the index comparison between two indexes composed of different combinations of indicators. This helps identify the areas that are more sensitive to specific patterns. Figure 4.3 represents the index comparison feature for the selection of two different indices. From the selection, pie chart visualization is displayed to show the composition of two different indices. The map interface shows the difference between those two indexes. The user click interaction on the geographic area will display the overlay bar chart

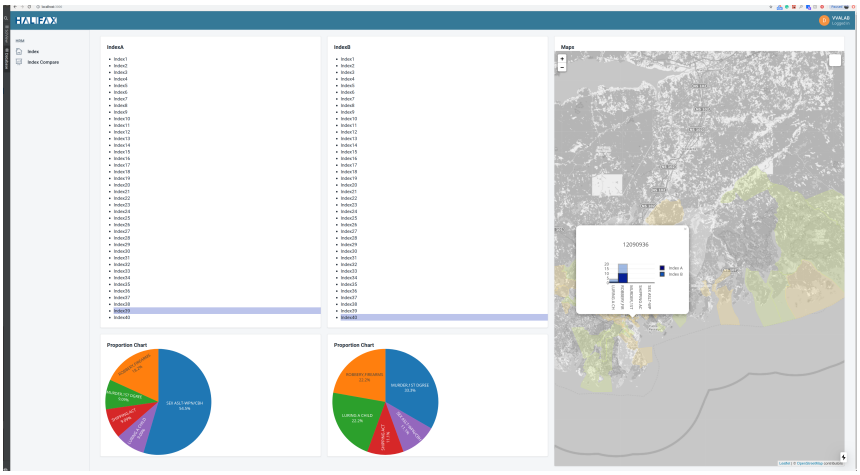


Figure 4.3: Illustration of Index Comparison feature for the selection of two different indexes. The pie chart represents the indicator composition for two indexes. The map shows the difference between Index B and Index A. The areas in the shade of green demonstrate the region with Index B having higher value and the areas in the shade of red indicates the the region of Index A having higher value. The overlay bar chart represents the indicator composition of Index A and Index B for the selected geographic area.

with the indicator composition information for two indices. It also helps narrow down the geographic area that has more concentration of different indexes. This flow of visualization makes it easier to interpret the composite index and stands unique on its approach than the existing visualization techniques for index [38, 44, 3].

The last usage scenario of our tool is the comparison of the different crime indexes generated for the years 2007 and 2008. Figure 4.4 represents the Comparison graph that shows the difference in index composition for the indexes generated for two different years. The pie chart define the indicator composition for the selection of two indexes. In the map, the areas that are non-white represent differences between Index A and Index B. Green areas indicate the region where the index for the year 2008 having higher values, and red areas indicate the regions where the index for the year 2007 is higher. Through this comparison, geographic regions with crime distribution growth or decline can be identified. This measure can also help in identifying the differences between two different indexes for the same year, and crime distribution spread for those dissemination areas. So, the dissemination area with a typical pattern of two crime indexes can be identified easily.

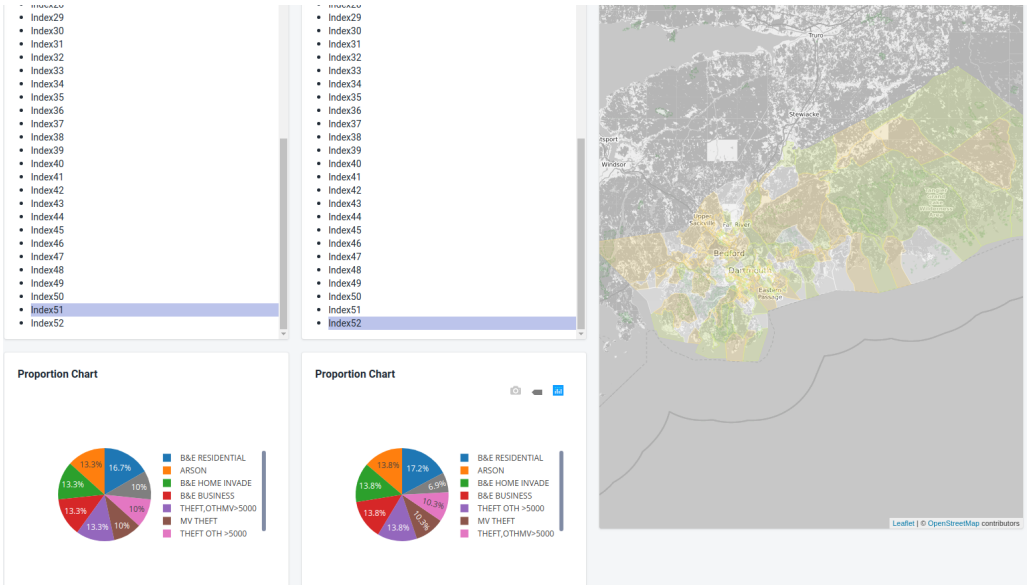


Figure 4.4: Illustration of Index Comparison feature for the selection of indexes generated for the year 2007 and 2008. The two pie chart shows the similar indicator composition for both indexes for different years. Green areas in the map indicates where 2008 index has higher impact than 2007 index. And Red areas represent the higher impact regions of 2007 index.

4.4 Index Interpretation

In general, index interpretation is performed using the indicators that are used to compose an index [44]. But, our proposed approach explores the relation between external indicators and a given index, allowing users to identify other aspects that can describe the index values. In this process, firstly, we employ regression analysis to approximate the external variables to the index using an AutoML tool called TPOT[7]. After that, the interpretation is conducted using the LIME strategy[36], which provides the impact of each indicator to the regression model.

As TPOT employs a genetic algorithm to identify the model that better represents the data relationships, it requires setting the number of population and number of generations. In our tests, we decided to set both parameters to 25, because TPOT presents a high processing time, and this value already present reasonable results. In this scenario, the regression took around 10 minutes to produce the results in a machine with Intel(R) Core(TM) i7-7820X CPU (3.60 GHz) and 64 GB of RAM.

Figure 4.5 shows an example of index interpretation. In this figure, the map

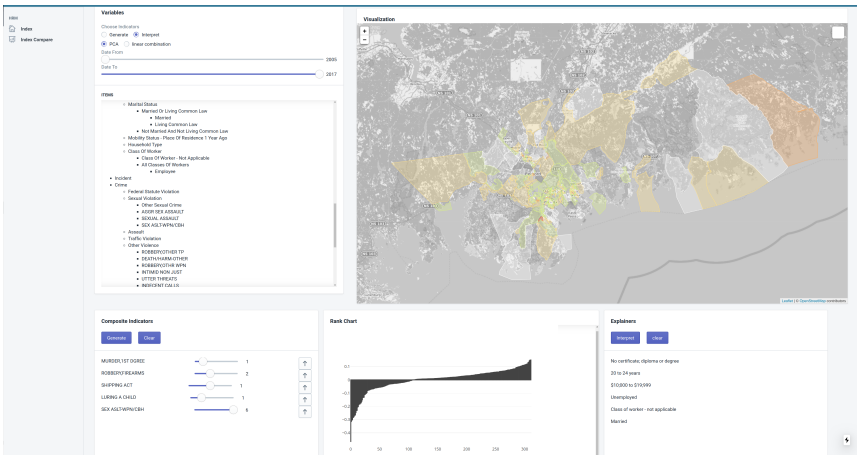


Figure 4.5: Illustration of the interpretation produced by the Approximation model for the composite index with the explainers *No certificate degree or diploma*, 20 *to* 24 *years*, *$$10000to*$$19999, *Unemployed*, *Class of Worker not applicable* and *Married*. The barchart represents the difference between predicted index and original composite index for all geographic areas.The values closer to zero indicates the perfect approximation relation.

is colored according to the difference between the original index and the prediction produced by the regression model. The lighter colors indicate a marginal difference, meaning that the association between the selected external indicators and the composite index created is high. Based on the color distribution, it is possible to identify the geographic areas where the selected explanatory variables best correlate with the given index. In our framework, users have the opportunity to choose the explanatory variable, allowing them to explore and analyze the data according to their knowledge. It is worth to mention that our framework uses an AutoML strategy to compute the regression model, avoiding the manual search for a good model.

The proportion of the explanatory variable to the approximation model produced for a specific geographic area is shown in Figure 4.6, in which LIME [36] is applied to explain the approximation model under analysis. Through this method, explanatory variables' contribution to a dissemination area can be visualized.The bar chart is visualized on top on each area to represent the contribution of explainers. The Rank chart represents the strength of the association between explainers and index. The values that are closer to zero indicate the stronger association.

To conclude, our approach enables us to identify the correlation of the explanatory variables with the composite index through regression analysis similar to the strategy

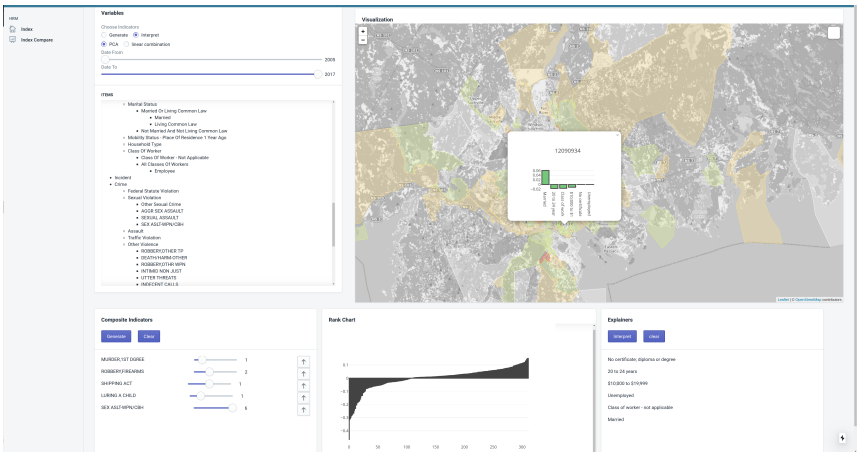


Figure 4.6: Illustration of explainers contribution from the interpretation model for the composite index.The user interaction on geographic area displays the popup with bar chart representing the explainer proportion. The regular bar represents the explainer with positive contribution and inverted bar indicates the explainer with negative contribution for the index.

used in literature [40], making the process convenient and easily accessible for end-users. Our framework helps to identify the specific geographic areas that are best described by the explanatory (external) variables and reduces the time needed to identify the right variables which makes the decision-making process more convenient.

4.5 Use Case: Property Crime Index

In this section, we present a use-case similar to the study conducted by Statistics Canada in 2001 to understand the crime density distribution of Halifax Regional Municipality [40]. Similar to this study, we create a Property crime index using the indicators *Arson*, *Motor vehicle Theft*, *Break and Enter Home Invasion*, *Break and Enter Business*, *Break and Enter Residential*, *Theft other than Motor Vehicles* $>$ 5000, *Theft other* $>$ 5000 and *Theft other* $<$ 5000. We employ our simple linear combination using the weights as 4, 3, 4, 4, 5, 3, 4, and 3, respectively. Figure 4.7 illustrates the variable selection for the index composition and the composed index for the 601 dissemination areas of the Halifax Regional Municipality.

After the index composition, the property crime distribution can be analyzed using the map visualization, which explains the spread of crime for the 601 dissemination

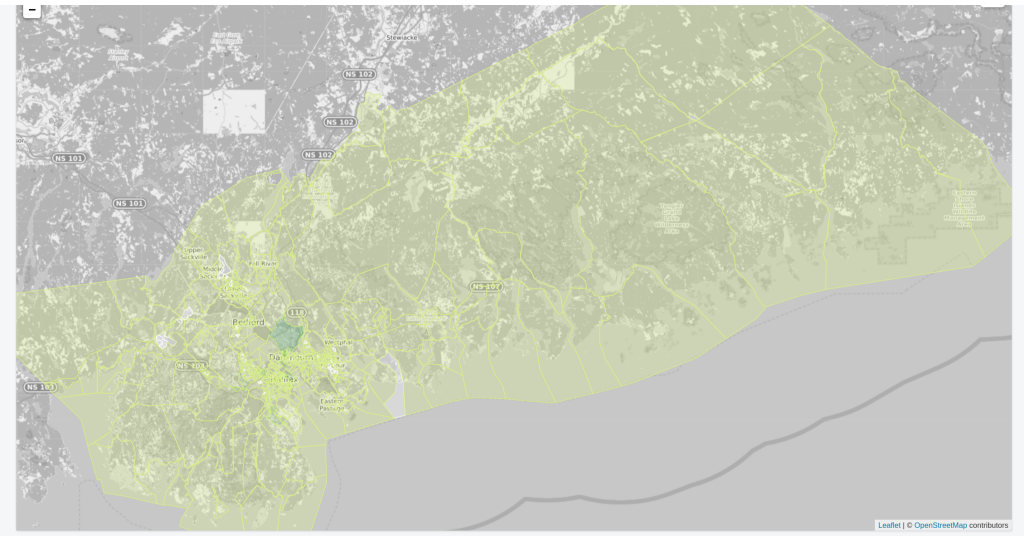


Figure 4.7: Illustration of Property Crime Index generated using indicators *Arson*, *Motor vehicle Theft*, *Break and Enter Home Invasion*, *Break and Enter Business*, *Break and Enter Residential*, *Theft other than Motor Vehicles* > 5000, *Theft other* > 5000 and *Theft other* < 5000 with weights 4, 3, 4, 4, 5, 3, 4, and 3 respectively.

areas of Halifax Regional Municipality. The bar visualization ranks the dissemination areas based on the index score, while the dissemination areas can be identified based on the hovering of the Rank Chart. Dissemination areas index composition is explained by the indicator proportion chart for each geographic area, in which figure 4.8 illustrates the indicator proportion for one dissemination area. In this figure, the indicator *Break and Enter Business* is the maximum impact indicator for the generated index, while indicator *Arson* is the minimum impact indicator.

Next, we compare the proposed index interpretation approach with the methodology adopted by the Statistics Canada [40], considering the explainers variables from the demographics of Statistics Canada of the year 2001. However, in our study, the explainers are from the demographics of the year 2016, which comprises the information from 2011 to 2016. Even so, the explainers variables used in this evaluation are similar to the study of Statistics Canada, which are *Median total income of households in 2015*, *Median after-tax income of one-person households in 2015*, *Percentage of Households spending more than 30 percentage on shelter costs*, *Population proportion that did not work*, *Not in the Labor Force*, *UnEmployed*, *No certificate; diploma or degree*, *High School Diploma certificate* and *University Certificate or diploma*.

In this scenario, we select the dissemination area with the higher crime index,

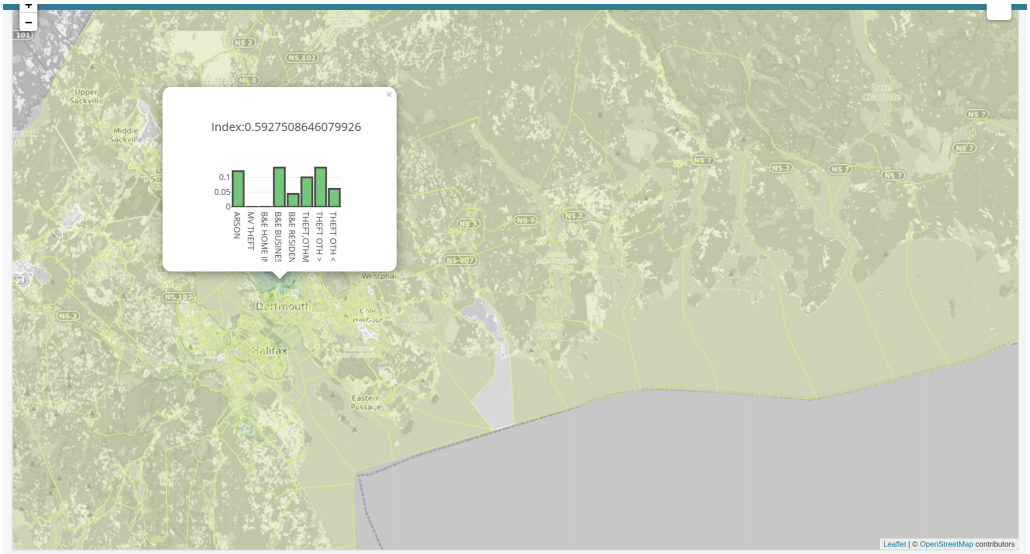


Figure 4.8: Illustration of the indicator contribution for the index composition of high property crime Index with value 0.5927.Dark green color pattern represents the area with high index.The indicator *Break and Enter Business* is the maximum impact indicator for the generated index, while indicator *Arson* is the minimum impact indicator.

equals to 0.59. The interpretation model for such area are correlated with the indicators *Secondary high school Diploma*, *Unemployed*, *Did not work in the work activity*, *No certificate or diploma*, and *Median Household income*. Besides, the only indicator with a negative correlation considering the same area is *percentage of owner households spending more than 30 percentage income on households*. Figure 4.9 illustrates the area with the higher crime index value and the indicators correlated to it.

Besides, we also evaluated the dissemination area with a low crime index in South Street, with the crime index equals to 0.07. For the interpretation, the positively correlated variables are *Median total income of households in 2015*, *Unemployed*, *Tenant households spending more than 30 percentage of income on shelter costs* and negatively correlated variables are *No certificate*, *Secondary High School Diploma* and *Owner households spending more than 30 percentage of income on shelter costs*. Figure 4.10 illustrates the area with the low crime index value and the indicators correlated to it.

For the high property crime areas, Statistics Canada found that “commercial zoning” and “unemployed” are the significant reasons for the high property crime rate neighborhoods. It agrees with some of The interpretation variables “unemployed”,

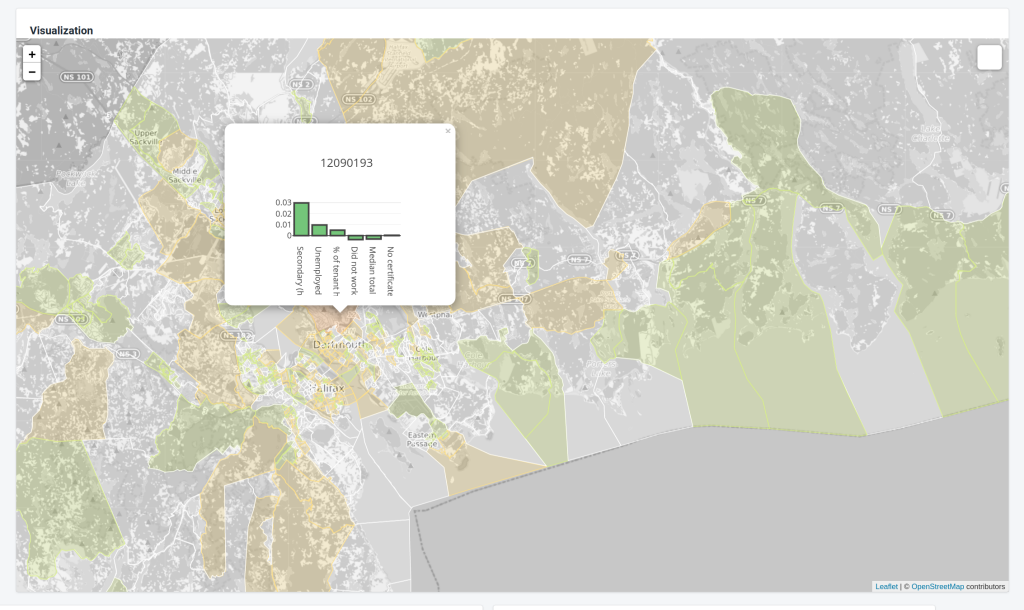


Figure 4.9: Illustration of the interpretation obtained using explainers. *Secondary high school Diploma*, *Unemployed*, *Did not work in the work activity*, *No certificate or diploma*, and *Median Household income*.The areas in white and lighter color represents the strong association between explainers and index. the Positively correlated variables are *Secondary high school Diploma*, *Unemployed*, *Did not work in the work activity*, *No certificate or diploma*, and *Median Household income*.

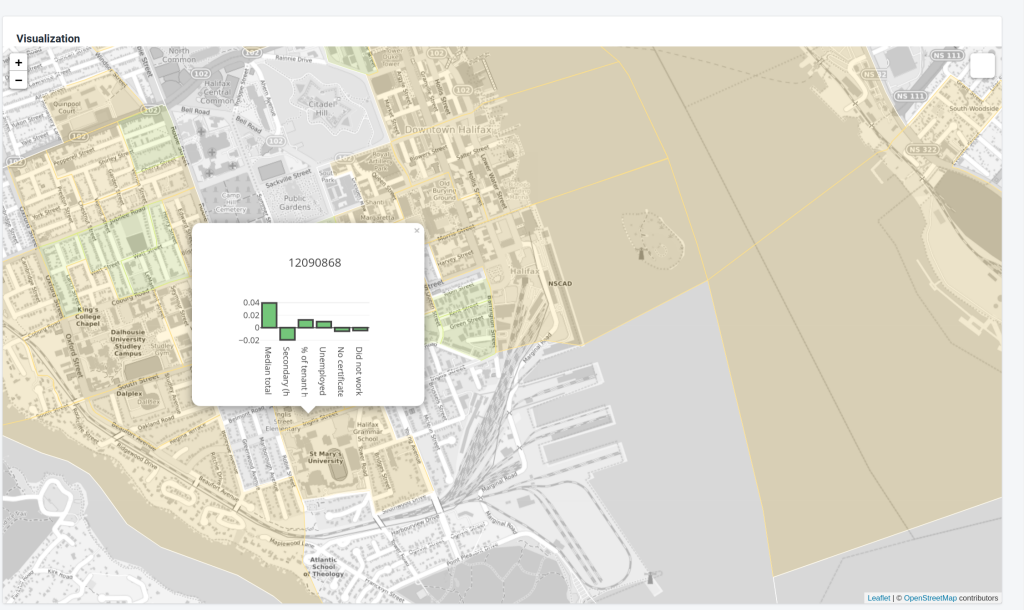


Figure 4.10: Illustration of interpretation obtained using explainers for the geographic area with low property crime index.The positively correlated explainers are *Median total income of households in 2015*, *Unemployed*, *Tenant households spending more than 30 percentage of income on shelter costs* and negatively correlated variables are *No certificate*, *Secondary High School Diploma* and *Owner households spending more than 30 percentage of income on shelter costs*.

"No certificate diploma or degree" produced by our tool for the most positively correlated variables for high crime index areas. Further, The statistics Canada [40] found that the variables "Median household income" and "proportion of population spending more than 30% of their income on shelter" correlates positively with the South-End region of Halifax. It matches with The indicators "Median total income of households", "unemployed and % of tenant spending 30%" produced by our system as positively correlated variables for the South-End region. Through these results, it is possible to observe that both high and low index areas and their interpretation results are similar to the observation produced by Statistics Canada [40], allowing us to conclude that our system is efficient enough to predict the appropriate external variables for the index interpretation. This agreement between our interpretation results and the findings provided by Statistics Canada indicates that our approach can successfully support the composition and interpretation of indexes.

4.6 Use-case: Violent Crime Index

For the second use-case scenario, we consider the indicators from the crime dataset that falls under the Violent crime category. The use-case simulates the study conducted by Statistics Canada to understand the high and low-density violent areas in Halifax for the year 2001 [40]. The indicators chosen for violent crime are *Assault Level* 1, *Assault Level* 3, *Assault Weapon/Bodily Harm*, *Murder* 1*st Degree*, *Murder* 2*nd Degree*, *Sexual Assualt*, *Robbery Firearms*, *Death/Harm-Other and Explosives*. We used the linear combination technique applying weights 3, 5, 4, 6, 7, 3, 3, 4, and 5, respectively.Figure 4.11 illustrates the variable selection for the index composition and the index distribution for the 601 dissemination areas of the Halifax Regional Municipality.

Figure 4.12 represents the indicator contribution for the high violent crime index area. *Aggr Assault Level 3*, *Assault Weapon/Bodily Harm*, *Murder 1st Degree* and *Assault Level1* are the list of indicators that contributed more for the high violent index.

To understand the violent crime index areas, we used the following explainers, *Bachelor's degree*, *Dwelling with Major repairs needed*, *Lone parent Family with female gender*, *Lone Parent family with children*, *Lone parent family with two children*, *20 to*

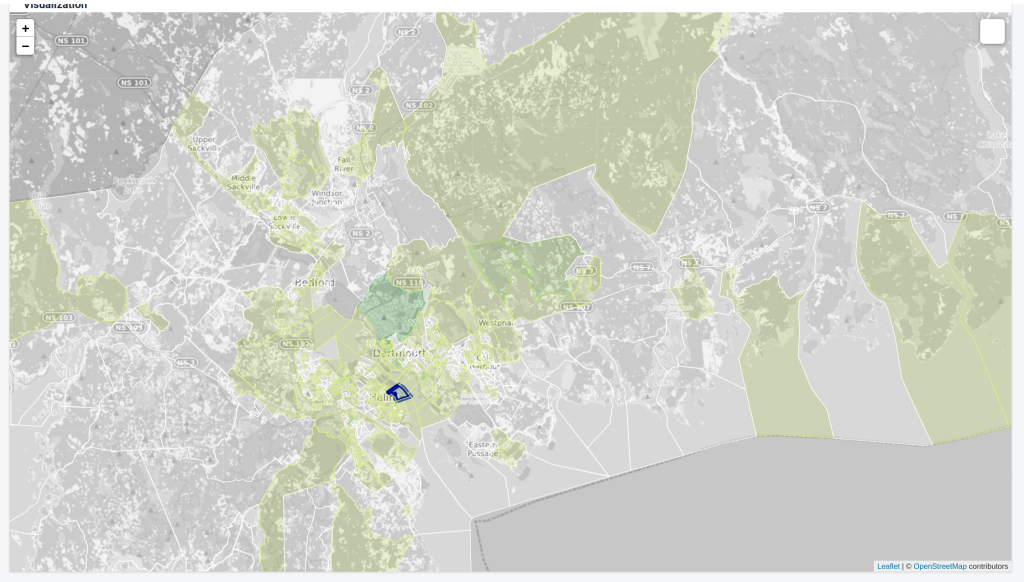


Figure 4.11: Illustration of Violent Crime index generated for 601 dissemination areas.The selected indicators are *Assault Level1*, *Aggr Assault Level 3*, *Assault Weapon/Bodily Harm*, *Murder 1st Degree*, *Murder 2nd Degree*, *Sexual Assualt*, *Robbery Firearms* and *Death/Harm-Other and Explosives* with weights 3,5,4,6,7,3,3,4and 5 respectively. The indicator proportion for crime index is represented for each geographic area.

Figure 4.12: Illustration of the indicator contribution for the index composition of high violent crime area among the 601 Dissemination areas of the Halifax Regional Municipality.The indicators with maximum impact are *Aggr Assault Level 3*, *Assault Weapon/Bodily Harm*, *Murder 1st Degree* and *Assault Level1*.The dark green indicates the areas with high violent index.

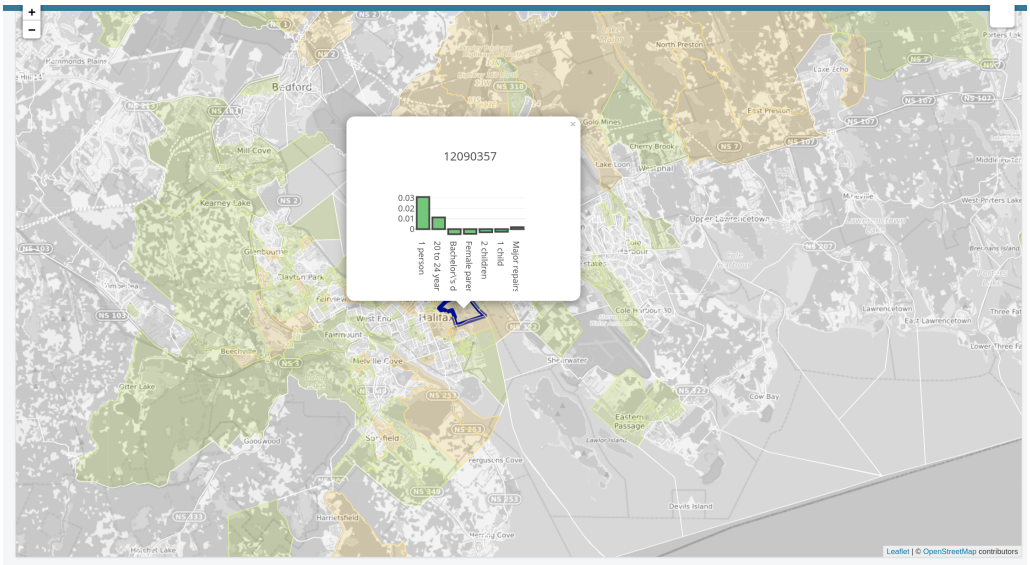


Figure 4.13: Illustration of interpretation obtained for the geographic area with high violent crime index.The positively correlated explainers are *1 person* and *20 to 24 years*. The negatively correlated explainers are *Bachelor's degree* and *Female Parent*. The white and light colored areas represent the stonger association between explainers and index.

24 years, and *Household size of one person* to interpret the generated index.

Figure 4.13 describes the explainers relation with the generated index for the Northeast region of the Halifax Regional Municipality. The selected region represents the high violent crime index. The explainers that are positively correlated with the index are *1 person*, *20 to 24 years*. The negatively correlated explainers for that area are *Bachelor's degree* and *Female Parent*. Statistics Canada identified the Northeast region of Halifax as high violent crime area. They found the proportion of the population with a bachelor's degree negatively correlated with the high violent crime for the North-End region. This finding matches with our system explanation for the high violent crime index area. The Bachelor's Degree is negatively correlated in both the Statistics Canada results and the results produced by our system.

Figure 4.14 represents the explainers for the violent crime index generated for the Southwest regions of the Halifax Regional Municipality. The most positively associated explainers are *Household size of 1 person*, *Dwelling with Major Repairs needed*, *Lone parent family with 1 child*, *Lone parent family with Female Parent*. For the Southwest region, Statistics Canada found that the *proportion of population living*

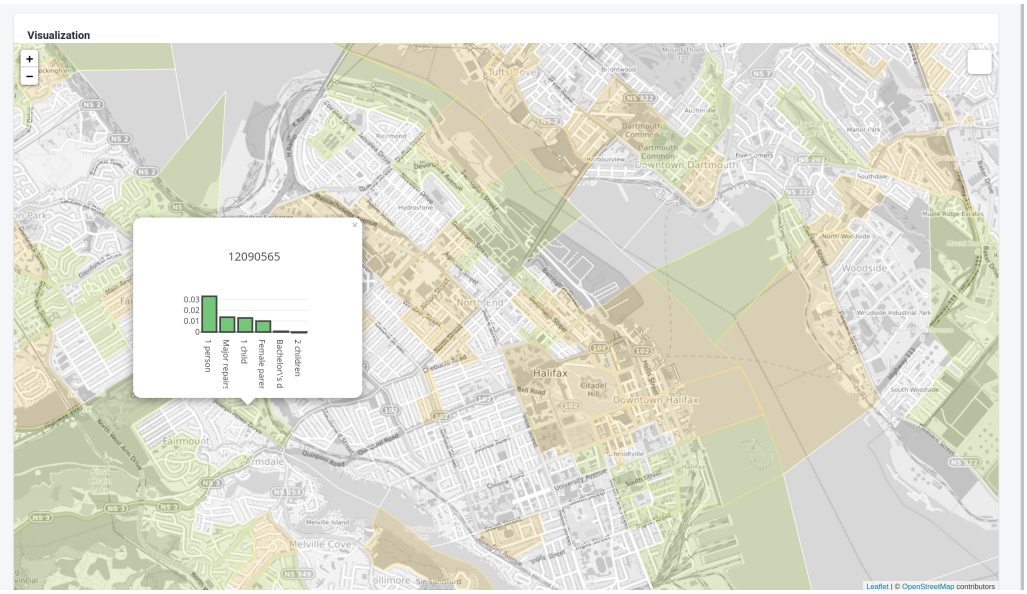


Figure 4.14: Illustration of interpretation obtained for the south west region of Halifax.The most positively associated explainers are *Household size of 1 person*, *Dwelling with Major Repairs needed*, *Lone parent family with 1 child*, *Lone parent family with Female Parent*. The light colored areas in the map represents the stronger association between index and explainers.

alone produced the largest positive contribution to the explanation of violent crimes for that region. The second positive contribution is the *lone parent mother family*. Finally, the *proportion of dwellings with major repair* also associated positively with the violent crime for the Southwest region. Therefore, Statistics Canada's observation matched with the interpretation produced by our system. Thus, the similar agreement between our results and statistics canada observations demonstrates the efficiency of our tool in interpreting the violent crime indexes generated for the Halifax Regional Municipality.

4.7 Use-case: Drug-related Crime Index

Initially, an index was created considering some indicators from HRPC dataset, which are related to drug crimes. In this scenario, categories selected were *p*roduction of heroin, *p*roduction of cocaine, *p*roduction of other drugs, *i*mportation of cannabis, *i*mportation of other drugs, *t*raffic of heroin, *t*raffic of cocaine, *t*raffic of cannabis, *t*raffic of other drugs, *t*raffic of crystal meth, *b*reak and enter other places, *h*uman traffic, *e*xplosives, *b*reak and enter firearms, and *i*nstruction of terrorism, having 5,

Figure 4.15: Illustration of indexes produced for the drug crime index with indicators *p*roduction of heroin, *p*roduction of cocaine, *p*roduction of other drugs, *i*mportation of cannabis, *i*mportation of other drugs, *t*raffic of heroin, *t*raffic of cocaine, *t*raffic of cannabis, *t*raffic of other drugs, *t*raffic of crystal meth, *b*reak and enter other places, *h*uman traffic, *e*xplosives, *b*reak and enter firearms, and *i*nstruction of terrorism having weights 5, 5, 4, 4, 4, 5, 5, 4, 4, 5, 5, 3, 5, 5, 6 respectively.The Color pattern in the map ranges from light green to dark green indicating the intenstiy of the index.

5, 4, 4, 4, 5, 5, 4, 4, 5, 5, 3, 5, 5, 6 as combination weights, respectively. Figure 4.15 shows the attained index. Using this visual representation, it is possible to analyze the impact of each indicator to the index in a specific dissemination area. For instance, Figure 4.16 presents the analyze of a particular area, in which the index value is 0.3529 and has *i*nstruction of terrorism as the most relevant indicator.

Next, the weights of indicators was modified to 4, 4, 3, 1, 3, 4, 4, 1, 3, 4, 5, 3, 3, 4, 6, respectively, in order to evaluate the impact of the weights in the composite indexes, as presented by Figure 4.17. In this example, we select the same dissemination area of the previous example for analysis, with value equals to 0.362. Figure 4.18 presents a comparison between both composite indexes, revealing regions where indicators affect the index value significantly. In this visualization, the map illustrates the impact in all dissimination areas, in which blue color means that the new index is lower than the original index, while the red color shows the opposite.

As a next step, explainers extracted from the HRMD dataset were selected to

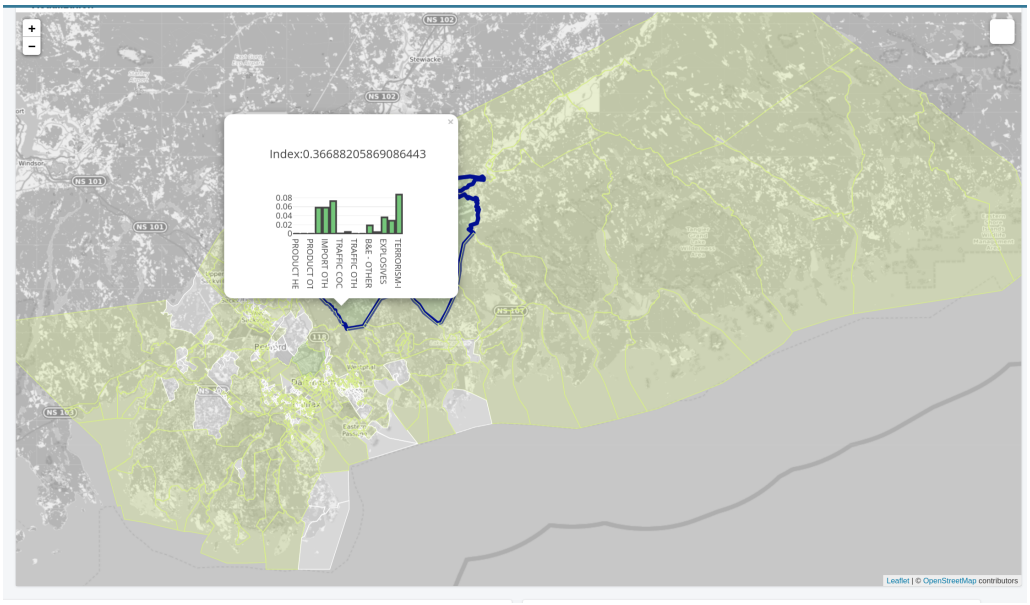


Figure 4.16: Illustration of an index produced for dissemination area with high crime index, having a value of approximately 0.368,*instruction of terrorism* is the most indicator with maximum impact for the generated index.

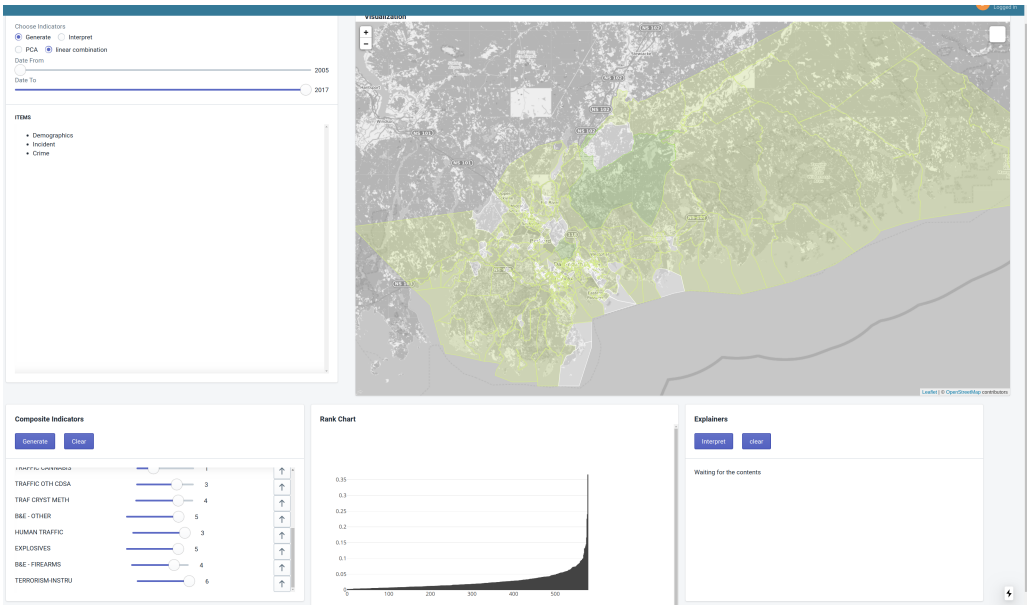


Figure 4.17: Illustration of indexes produced for the drug crime index with indicators *p*roduction of heroin, *p*roduction of cocaine, *p*roduction of other drugs, *i*mportation of cannabis, *i*mportation of other drugs, *t*raffic of heroin, *t*raffic of cocaine, *t*raffic of cannabis, *t*raffic of other drugs, *t*raffic of crystal meth, *b*reak and enter other places, *h*uman traffic, *e*xplosives, *b*reak and enter firearms, and *i*nstruction of terrorism having different weights 4, 4, 3, 1, 3, 4, 4, 1, 3, 4, 5, 3, 3, 4, 6, respectively.

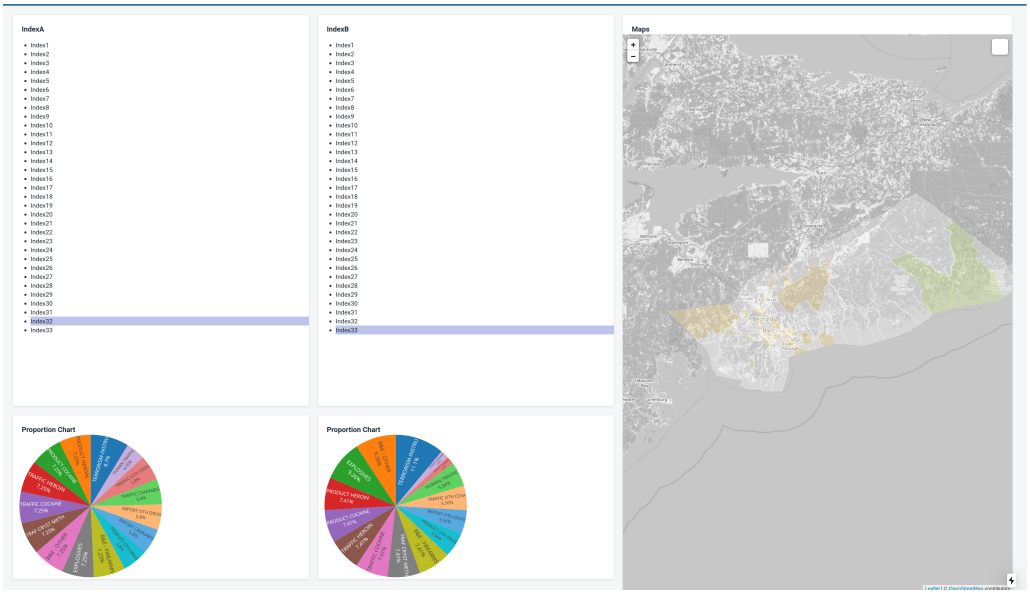


Figure 4.18: Illustration of index comparison. In the up-left corner of the screen, the menu of the index is presented, a pie chart on the bottom left shows the percentage of each indicator used to produce the composite index, and the map on the right shows the difference between them. Green color means that the index on the right is higher than the index on the right, while the red color shows the opposite.

interpret the first index. Demographic information such as income, householder, and civil status was used to evaluate if these characteristics are related to the drug crime index.The Explainers selected are *Canadian citizens aged 18 and over*,*Household size with 5 or more persons*, *Low income population with 18 to 64 years*,*population that are never married*, *Occupation named Natural and applied sciences and related occupations*, *Population that are married*, *Household size with 3 persons*, *After tax income under* 10000, *Without after tax income*, *After tax income with* 20000 *to* 29999, *Age characteristics under* 25 *to* 29 *years* and *Age characteristics under* 30 *to* 34 *years*. Figure 4.19 shows the results of such index interpretation, providing the reliability of these explainers, indicating the difference between the original indexes and the predicted indexes. The values that are closer to zero in the Rank chart indicates the better approximation achieved by the model explainer to original index.

Figure 4.20 presents the explanation obtained for a dissemination area with high crime index. In this figure, the variables that are more positively related to the index in the area are *Low income in 2015, 18 to 64 years* and *proportion of population that is never married*. The most negatively correlated explainers are *Canadian citizens aged*

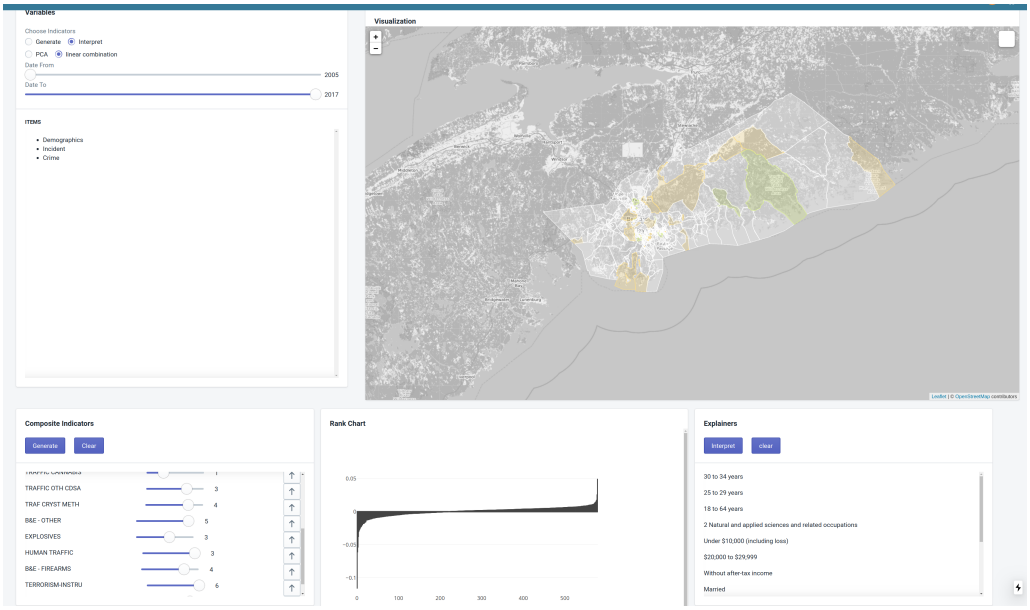


Figure 4.19: Illustration of the explanation obtained for the composite index using the explainers *Canadian citizens aged 18 and over*, *Household size with 5 or more persons*, *Low income population with 18 to 64 years*, *population that are never married*, *Occupation named Natural and applied sciences and related occupations*, *Population that are married*, *Household size with 3 persons*, *After tax income under* 10000, *Without after tax income*, *After tax income with* 20000 *to* 29999, *Age characteristics under* 25 *to* 29 *years* and *Age characteristics under* 30 *to* 34 *years*. Light colors on the map represent high reliability, while dark colors correspond to low reliability.

Figure 4.20: Illustration of the explanation obtained for a specific dissemination area which has high index value and a reliable explanation. The variable that is most positively correlated to the composite index in that area is *Low income in 2015 18 to 64 years* and the most negatively associate explainer is *Canadian citizens aged* 18 *and over*.

18 and over and *Household size with 5 or more persons*. This concludes the reliable explanation obtained for the high drug and violent crime index of that dissemination area. This analysis helps the expert to identify the potential reason behind the specific crime happening for the dissemination area.

The above three scenarios illustrate the use case of our tool in the index composition and interpretation. As observed in the first two scenarios, the tool has been effective in the composition of the index and produced efficient model interpretation for the explainers. The interactive user interface reduces the burden of choosing the right explainer for identifying the reason behind the generated index. The intuitive nature of the user interface makes the tools accessible for end users and help them to understand their generated index without the aid of the expert. This makes the tool powerful and with initial training for the tool, end users will be comfortable with the complete flow of the user interface. This concludes the motive of the tool to be an powerful interactive user interface for index composition, which could help in the process of making better decisions.

Chapter 5

User Study

The initial requirements for designing the interface has been gathered in the early stages of development of interface. To determine the features required to accomplish the task, we had group meetings with the people from Halifax Regional Municipality and Halifax Regional Police. There were periodical discussion with them during the development phase to identify and evaluate the visual interface to make sure it matches the expectations. Their suggestions in those meetings helped us to identify the appropriate feature for designing the interface. Further, they provided valuable comments to improve the comprehensive nature of the interface. This helped us to design the accessible visual interface that helps the end users to easily navigate across the features built for Index Composition and Interpretation.

An exploratory user study was conducted to evaluate the performance of our tool and determine whether it is comprehensible for the users. The assessment was designed with the intent to determine whether the users could accomplish specific tasks. Through task accomplishment, the level of agreement among different users is computed to determine the accessibility of the tool. The assessment had further components to gather feedback from the users on the interface design and software quality.

Fifteen participants took part in the study. Eleven participants are male, and 4 participants are female, comprising 10 undergrad students and 5 masters students from the computer science program at Dalhousie University. All the students were encountering the tool for the first time, though they had an extensive background in Data Visualization and Machine Learning (13 moderate to high familiarity and 2 moderate familiarity). Seven individuals rated their familiarity with composite index high, and the other 8 individuals declared their familiarity moderate to low. Although the participants had a similar computer science background, they represented various skills regarding the familiarity with map-based interactive user interfaces and the

understanding of dissemination areas.

The study experiments were conducted in a remote procedure. The participants were given a demonstration video of the tool for 10 minutes. Each individual has an opportunity to play around with the tool and ask questions after the demonstration. After that, participants were asked to fill the demographic questionnaire. The tasks were designed to let users use the tool's components and interactive functions in practice, for example, navigating through the indicators to make a selection, setting indicator weights, using the map interface, rank charts, and the index comparison feature. After completing the questionnaire, the participants were directed to take the session, which was composed of three parts. For the first part, they were asked to create a composite index given a set of indicators. They were asked to answer questions based on their interaction with the map interface for the generated index. For the second part, the participants were asked to select specific explainers to interpret the generated index. The tasks were structured to coerce participants answering questions for their interaction with the explainer output on the map interface. For the last part, the participants were asked to create another index by changing the weights of the selected indicators for the previous stage. They were directed to answer questions for their interaction with the comparison feature for the two indices. The questions are ordered from easy to difficult tasks. This helped participants to build their skills as they progress in the study. After the end of the task, the participants were asked to respond to 5 point Likert scale questions about their subjective opinion on the functionality, quality, and usability of the interface. Each study session lasted for 35-50 minutes.

5.1 Study Results

Fourteen different questions comprise the study. The questions are structured so that the first 2 questions cover the index generation task, the next 6 questions are related to the index interpretation task, and the last 6 are related to the index comparison feature.The descriptive statistics is computed for the participants response and it is presented as follows. For the first two sections of the questions, the participants are asked to identify the distinct locations in the map, specifically the North-End and South-End regions of Halifax Regional Municipality. The participants were asked to

identify the indicator with maximum and minimum impact for those two regions. 100% of the participants responded with the indicator *Explosives* for the South-End region and 86.7% responded similarly for the North-End region.

For the next section of questions, the participants were asked to interpret the generated index by selecting specific explainers from demographics indicators. They were asked to identify the same location in the map used for the index generation task and identify the variables that positively and negatively impact those regions. The participants responded with 100% for the determination of a positively associated explainer for the region. For the negatively associated explainers for the South-End region, participants responded similarly with the rate of 93%. The participants were then asked to identify the positively associated explainer for the highest and lowest index regions. They responded similarly with the score of 80% for the interpretation of the highest index area, and 66% for the interpretation of lowest index regions. Since the AutoML configuration was minimized to facilitate the study faster, the results obtained for the interpretation stage of some study varied slightly. That might explains the reason for lower agreement among the users for the lowest index regions.

For the last stage of this section of questions, the participants were asked to identify the impact of adding/removing indicators to the generated index. They were instructed to remove a specific indicator called *Explosives* for generating the new index. They were asked to identify the impact of this removal operation on the North-End region. They responded similarly with 100% for the impact identification task. The participants responded similarly with 93% for the identification of the impact of adding a new indicator *B&E Home Invade* to the previous index.

The final section of questions asks the participants to generate new indices with a set of given weights to be applied to the previously chosen indicators. They were asked to select the new index and the previous index for comparison, here called Index B and Index A, respectively. They responded with 100% for the recognition of indicator composition for the selection of both indices. Next, the participants are instructed to identify the region near Downtown Halifax where Index B has higher precedence over Index A. 93% of participants recognize the same area. The participants were

asked to identify the indicator composition for the two indices for a directed dissemination area in the map interface. They were asked to choose the indicator with the maximum difference between Index A and Index B values. 100% of the participants responded similarly with the selection of the appropriate indicators. The final tasks are directed towards the opinion-based question with the statement of Index A having higher/lower values than Index B for the dissemination area. 100% of the participants agreed with the statement of Index A having higher values than Index B, and 93.3% of the participants disagreed with the statement of Index A having a lower impact over Index B.

The agreement of users for all the questions is calculated using standard statistical analysis for the inter-rater agreement among multiple users. We consider two statistical analysis measures for multi-rater based agreement models, Kripendorff alpha measure [27] and Fliess Kappa Coefficient [15]. The user responses for all the questions are encoded for all before applying those statistical measures.

After the application of those measures for the 15 users and 14 questions, we obtained the Krippendorff score of 0.8534, that is, a perfect agreement according to the interval of Krippendorff alpha[27]. And, we obtain the Fliess Kappa coefficient score of 0.8527, a perfect agreement score among multiple raters as per the interval of Fleiss Kappa [15].

After the completion of the study, users responded to close-ended questions to evaluate the functionality, user interface design, and usability of the tool. The data was collected using a 5 point Likert scale for all the post-evaluation questions.

The table 5.1 presents the functionality questions and a summary of the answers submitted by the users. Most participants agreed with the non-complicated functionality of index creation. Unexpectedly, only two participants agreed with the complicated process for generating indices. Furthermore, most participants confirmed it is easier for them to identify the geographic areas with higher/lower indices through the bar chart visualization functionality. The participants disagreed with the complexity of the interpreting feature.

Regarding index comparison, the pie chart feature was found to be useful for comparing indexes (53.3% strongly agree, and 33.3% agree). Almost all of them agree with the intuitive nature of the index comparison feature (40% strongly agree,

Table 5.1: Post Evaluation Questionnaire review of the 15 users for the functionality of the interface.

Question	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat Agree	Strongly Agree
Is it complicated to create Index	6	4	3	0	2
Is it complicated to identify the dissemination area with higher and lower index in map from the rank bar chart?	8	4	1	1	1
I find it challenging to use interpretability functionality.	5	6	3	1	0
Index Comparison feature does not look that intuitive to me.	6	5	2	1	1
Pie chart feature makes it easy for me to compare the index in a better way.	0	2	0	5	8
Overall, I find the design of the system interesting and accessible.	0	0	1	8	6

and 33.3% agree). Finally, almost all participants agree with the functionality of the visual interface intriguing and more accessible (40% strongly agree, and 53.3% agree).

Table 5.2 summarizes the participants' ratings for the user interface design. The diverse response is obtained for the comprehensive nature of the map interface used in the system. This could be improved by restricting the navigation of the map interface to only geographic areas relevant to the dataset. Only a few participants have disagreed with the color pattern used for the index visualization (6.7%, strongly disagree, and 13.3% disagree). Almost all participants agreed with the click functionality of the map to identify the indicator/explainer composition more accessible (40% strongly agree, and 46.7 % agree). They also found that switching map layers between Index and Interpret layers more accessible (53.3% strongly agree, and 26.7% agree) and intuitive to comprehend the index (53.3% strongly agree, and 40% agree). Participants agreed on the intuitive nature of visualization (53.3% strongly agree, and 40% agree) and found rank chart visualization to be more convenient to identify

the high index geographic areas in the map (53.3% strongly agree, and 40% agree). Most participants agree with the statement it is easy to navigate to create an index (40% strongly agree, and 46.7% agree) and switching to explainers to interpret the generated index (46.7% strongly agree, and 40% agree). They also mostly agreed with the purpose of index comparison features to make the understanding of different indices easier (53.3% strongly agree, and 33.3% agree).

Table 5.3 presents the rating submitted by the users for the usability of the interface. Most participants did not find the system to be unnecessarily complex(Fir System complexity, 26.7% strongly disagree and 60% strongly agree). There is a diverse response from the participants for the statements that the users needed to learn more before accessing the system, and most users would learn to use the system easily. These inconsistencies might have been caused by the participants' diverse expertise and background. Most participants agreed that the system navigation is accessible(33.3% strongly agree and 53.3% agree) and felt comfortable in accessing all the features of the system(40% strongly agree and 40% agree). They were also satisfied with the consistency maintained by the system for all processes(For system inconsistency 40% strongly disagree and 60% agree).Thus the final study results was satisfactory that most users find the user interface accessible and effective in regards to the visualization of index composition and interpretation.

Table 5.2: Post Evaluation user review of 15 users for the User interface Design of the tool.

Question	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat agree	Strongly Agree
I prefer the map interface of the system to be more comprehensible.	2	2	6	2	3
I like the color scale used for the Map interface.	1	2	3	5	4
I find it straightforward to switch between layers in the map interface	0	2	1	4	8
The visualization-based system is intuitive to use.	0	0	1	6	8
Map-based UI with different layers is so intuitive to comprehend the index.	0	0	1	6	8
Rank chart makes it easy to identify the dissemination areas with higher index score	0	0	2	5	8
Index comparison feature makes the process of understanding index easier.	0	1	1	8	5
I find it easier to navigate through the indicators to create the index.	0	0	2	7	6
I find it straightforward to switch to explainers to interpret the generated index.	0	1	1	6	7
I find it easier to identify the indicator proportion on the click functionality of the map interface.	0	0	2	7	6

Table 5.3: Post Evaluation user review of 15 users for the Software Usability of the system.

Question	Strongly Disagree	Somewhat Disagree	Neutral	Somewhat agree	Strongly Agree
I found the system unnecessarily complex.	4	9	1	0	1
I needed to learn a lot of things before I could get going with this system.	4	2	4	3	2
I would imagine that most people would learn to use this system very easily.	0	3	2	8	2
I find that overall navigation of the system is quite accessible.	0	0	2	8	5
I felt so comfortable in accessing all the features of the system.	0	0	3	6	6
I thought there was too much inconsistency in the system.	6	9	0	0	0
I thought I would need the support of the technical person to be able to use the system.	4	5	2	3	1

Chapter 6

Conclusion

Composite Index Generation is widely employed to provide a representation of an abstract concept being a composition of a set of indicators related to a geographic area. Such composition faces various challenges, including the importance of specific indicators and different perspectives to interpret it. In this context, we propose an approach that aims to improve the interpretation of an index composition using a visual analytics strategy, allowing the user to conduct comparisons between different pairs of indices and associating external variables (explainers) for interpretation to an index under analysis. In this context, the proposed framework uses information from the Halifax Regional Municipality, which contains information about 601 dissemination areas, including demographics and crimes.

In our system, we decided to devise a simple linear combination strategy to compose an index, instead of using PCA given two main reasons. First, we aim to address situations where an index does not match the user expectation since, in our case, we allow users to define the combination weights while PCA uses data variance. Second, this also avoids the errors produced when uncorrelated indicators are used. It is worth mentioning that the selected weights can be positive or negative, something that is also not supported by PCA. In our system, once an index is created, it is presented using a map metaphor, where users can interact to see the contribution of each indicator for each dissemination area.

After index creation, interpretation tasks can be performed using regression analysis. In this case, the map visualization shows the correlation of explainers variables with the generated composite index. Also, the user can select an area in the map to see the relationship of each explainer with the index for a specific region. Moreover, different indices can be compared through a pie chart, indicating the contribution of each indicator. In summary, the proposed approach permits to produce and explore composite indices according to the importance of indicators defined by the user,

providing an interpretation of indices with support of external variables. Such exploration helps in the identification of findings or reasons for the obtained index values for specific regions considering information not contained in the data used to generate the index.

The system is evaluated using the baseline scenarios presented by Statistics Canada to understand the high crime density areas of Halifax Regional Municipality for the year 2001 [40]. They conducted the study to understand the violent and property crime regions of Halifax. They applied a simple linear regression model for selected demographic variables to identify the reason behind the density of crime happening in Halifax. We simulated their experiment by generating indexes for property and violent crime. The explainers predicted by our system for index interpretation is like the results produced by Statistics Canada for the high-density crime areas. This validated our system capability to interpret any indexes without the expert intervention for identifying the appropriate machine learning model and feature selection. We presented one more scenario demonstrating the usability of our system in real-time to generate and interpret the indexes that represent the severe crime related to drugs.

A formal exploratory user study was conducted to assess the accessibility of the interface and to understand if the visual interface incorporates necessary features so that the most uses agree to with a similar conclusion for the index composition and interpretation. The users are guided with a scenario for their interaction with the user interface and are asked to answer questions from their interaction. Two standard Statistical Levels of Agreement, viz., Krippendorff's alpha, and Fleiss Kappa, were applied to measure the level of agreement between different users. The study outcome was satisfactory since a perfect agreement is achieved among different users according to the interval defined by the two statistics measures with the score of 86% for the Krippendorff's alpha and 85% for Fleiss Kappa. Most importantly, almost all users gave positive feedback regarding the usability of our Visual Interface. Users have stated that our tool is easier to navigate and that there is no inconsistency observed in its logic flow. They further identified the tool to be useful for generating and comprehending the index. The overall functionality of the interface was identified as intriguing and accessible by the users. The interface design and usability were rated positively by most users. Finally, the interface was found to be useful and easier to

learn so that users can get along with all the functionality of the system for the purpose of an index creation and interpretation. This concludes our purpose of designing an accessible interactive user interface for composing and interpret composite indexes using visual analytics and interpretability strategies.

6.1 Future work

Besides the positive feedback from the users, we have observed some limitations of the user interface that could have been improved to make the visual interface more accessible. And more importantly, many users have provided some valuable suggestions regarding the possible improvements to our interface. Further, we have identified some limitation regarding the interpretation technique for explainers. In this section, some future directions of the research are discussed.

First we found that some users had felt the need for the map interface to be restricted only to the area under the study. Therefore, in the future iteration of implementation, the map interface could be improved by restricting the accessibility of the location as per the boundary requirements of the loaded dataset. The next improvement could be made in the direction of indicator selection for composing indexes. The feature, that helps the user to choose indicator from the list of indicators, was recommended to have adaptability to select multiple indicators with minimum interaction. It can be improved in the future iteration of work. Third improvement could be made in the direction of adopting the best strategy for functionality or feature distinction. So the users could view only the appropriate visual components for their selection of operation like index composition or index interpretation from explainers. The same improvement could be applied for the area selection in the map interface. So the users could view only the interested geographic area in the map interface.

Selecting the appropriate indicators from the huge collection of data sets might require some easier procedure to identify the specific indicator. This improvement could be achieved by bringing in the new functionality related to the search bar to restrict the loaded indicators as per the query. This search bar with filter option could also improve the accessibility factor of the tool for the general user with limited knowledge about the organization of indicator hierarchy structure. In the current

interface, the user could compare their index with the previously generated indexes. But the user interface could have some additional option for the user to view the most commonly selected indicator for each category of the dataset from the previous user. The pattern recognition algorithm can be adopted to identify the indicators that are preferred often by the users for each category or sub category across the time for index composition. Similarly, for interpretation the similar pattern recognition could be applied to find the most preferred explainer selection for the specific combination of indicators.

Our final limitation of the system that could be improved is in the process of identifying the appropriate Machine Learning model for the purpose of Index Interpretation. With the current implementation, the perfect model generation from Auto ML technique is excellent in terms of finding the approximation relation between the dependent variable and the selection of independent variables, also called explainers. But, it takes more time for the task to identify the correct model. In the future, we could improve this technique by finding the correct strategies to generate the perfect approximation model with lesser time. With lesser time and better approximation, it will improve the usability of the visual interface and help users to reach conclusions at a faster rate.

The Endeavor of Using Visual Analytics and Interpretability strategies to understand the impact of input variables on Indices derived from Municipality Data Sets has received some positive and encouraging responses from the participants. Those comments appreciated our system for the interesting visualization system, balanced working of internal algorithms, user-friendly interface, and well-designed interactive and accessible interface. Although the thesis part is completed, the interface would be available for the real world application with next releases and improvements.

Appendix A

Ethics

A.1 Recruitment Document

Subject: Invitation to participate in a Research Project for Visual Interface

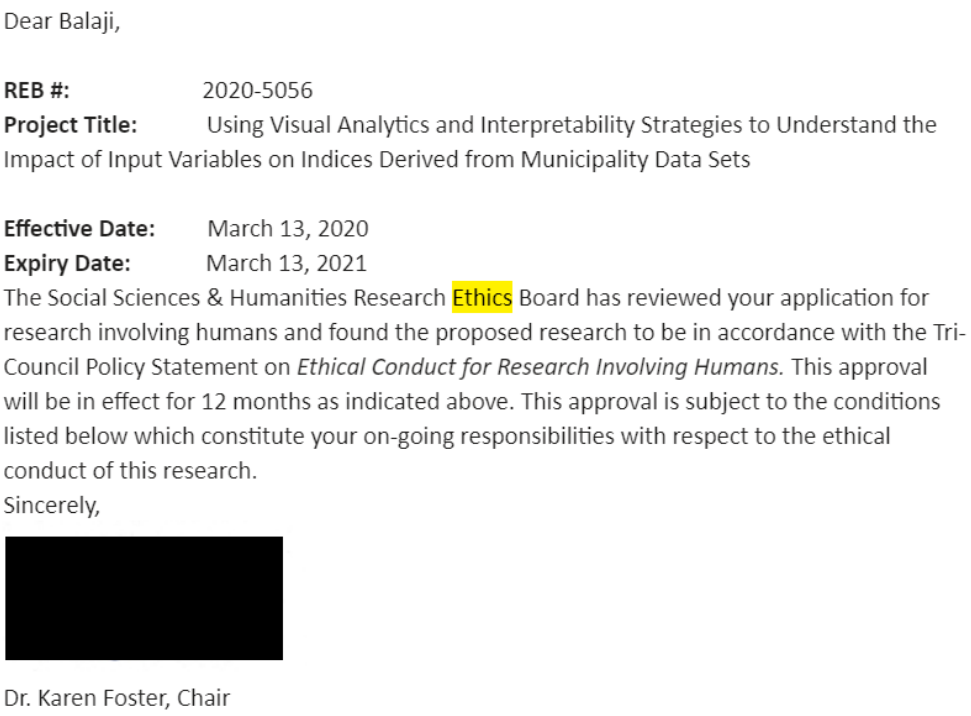
Hello Everyone,

We are recruiting participants to take part in a research study to evaluate the visual analytics application. The study involves the interaction with the Visual Interface tool designed to interpret the pattern of index generated for the Halifax Regional Municipality. We are looking for students/staff/professors at Dalhousie University. The research study is one session. The study will be conducted online through Microsoft Teams Platform. You will meet the lead researcher online who will elaborate on the study and ask you to send the signed consent form through email and fill in a background questionnaire. After that, you will have a short demonstration video of interacting with the user interface for about 5-10 minutes. After the completion of the study, you will fill out and submit an evaluation questionnaire about the experience with the system. You should be able to complete the study in 50 minutes. Compensation is CAD 15 for participation in the study. The study will be expired in a week from today. If you are interested in participating, please contact Balaji Dhakshinamoorthy (balajid@dal.ca).

The link to the consent form is attached to this email. If interested, Participants are recommended to read the consent form, sign upon the agreement of conditions and email to the principal investigator Balaji Dhakshinamoorthy (balajid@dal.ca) before the beginning of the Study.

A.2 Ethics Approval

The user study of this dissertation requires the approval from Research Ethics Board. The application process takes about 2 months that includes review of the submitted



Dear Balaji,

REB #: 2020-5056
Project Title: Using Visual Analytics and Interpretability Strategies to Understand the Impact of Input Variables on Indices Derived from Municipality Data Sets

Effective Date: March 13, 2020
Expiry Date: March 13, 2021

The Social Sciences & Humanities Research Ethics Board has reviewed your application for research involving humans and found the proposed research to be in accordance with the Tri-Council Policy Statement on *Ethical Conduct for Research Involving Humans*. This approval will be in effect for 12 months as indicated above. This approval is subject to the conditions listed below which constitute your on-going responsibilities with respect to the ethical conduct of this research.

Sincerely,

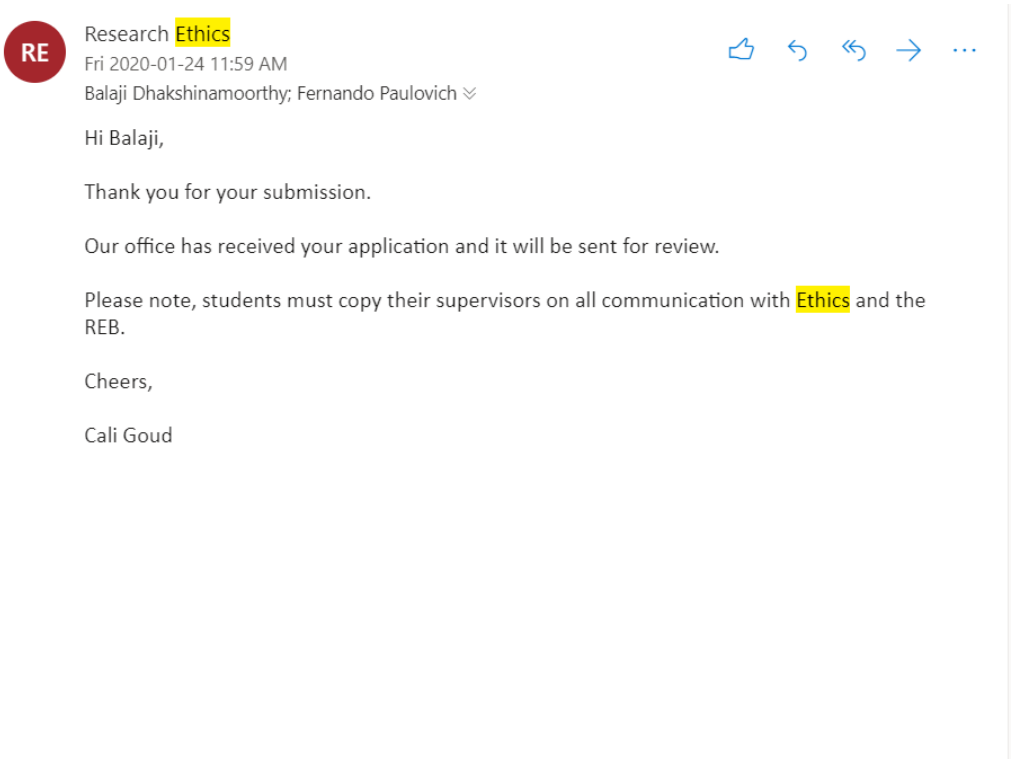
[redacted]

Dr. Karen Foster, Chair

Figure A.1: Ethics Approval Email

application. Once approved, the user study can be initiated with the condition listed in the application. This section displays the application submission and approval obtained from REB. Further, the details of the consent form and questionnaire are attached for the reference. FigureA.1 indicates the approval obtained from the REB for user study. Figure A.2 illustrates the application submitted for the REB.

Attached are the screenshots of the email conversation that describes the submission of application and Approval of the REB Board.



RE

Research Ethics

Fri 2020-01-24 11:59 AM

Balaji Dhakshinamoorthy; Fernando Paulovich

Hi Balaji,

Thank you for your submission.

Our office has received your application and it will be sent for review.

Please note, students must copy their supervisors on all communication with Ethics and the REB.

Cheers,

Cali Goud

Figure A.2: Email indicating the submission of Ethics Application

Appendix B

Consent Form

Using Visual Analytics and Interpretability strategies to understand the impact of input variables on Indices derived from Municipality Data Sets.

Principal Investigator: Balaji Dhaskhinamoorthy balajid@dal.ca(902-430-7106)

Supervisor: Dr. Fernando Paulovich (paulovich@dal.ca) (902-494-1986)

Faculty of Computer Science, Dalhousie University, 6050 University Ave., PO Box 15000, Halifax, NS, B3H 4R2, Canada.

Contact: balajid@dal.ca(902-430-7106)

You are invited to participate in a research study being conducted by Balaji Dhakshinamoorthy who is a Master of Computer Science student at the Faculty of Computer Science, Dalhousie University. This study is being done as part of a research project.

The information in this consent form will outline any possible risks, inconvenience and discomfort that you might experience.

Participation is voluntary, and participants are free to withdraw at any time without repercussions. If you have any concerns and questions about the study, please do not hesitate to ask the principal investigator.

Purpose The study is an exploration to determine the accessibility of the Visual Interface and to evaluate its interpretability among different users.

Study Design The principal investigator will analyze the operations the participants use during the experiment in order to design a better interface to aid decision making process. The principal investigator will also analyze the participant's final selected results.

Who can Participate in the study You may participate in this study if you are a Dalhousie University student/staff/faculty. You should be able to play around with interfaces like Map based UI (Google Maps), bar chart and pie chart. The study will evaluate the interpretation of the visual interface.

Who will conduct the research Balaji Dhakshinamoorthy, an MCS student from the Faculty of Computer Science, Dalhousie University, will be conducting the research, i.e., distributing and collecting questionnaires, introducing and demonstrating the software, and answering questions.

Possible risks and discomforts No extraordinary risks are anticipated in the present study. The risk is expected to be no more than that of every day's life. The prolonged duration of 50 minutes might make you little exhausted. But there will be breaks in between sessions. Your name will not be connected to the data collected from you.

Possible benefits You will be given $15 CAD for compensation. In addition, your participation will be greatly appreciated, and we expect that it will help us to learn better how users can interact with the Composite Index Generator systems and develop new technique to aid people in understanding the complex data better.

What you will be asked for You will answer a demographic questionnaire, receive a short training on the task and systems used in the study, then you will perform three tasks and answer the questionnaires about the tasks you performed. The study requires Participation time of 50 minutes which includes the system demonstration time of 5-10 minutes.

- You will sign the consent form.
- You will complete a demographic questionnaire.
- You will be given a tutorial on how to use the Interface.
- You will be given a practice session to use the software.
- You will be given evaluation (post-condition) questionnaire.
- You will perform three tasks of generating Indices, interpreting indices and comparing indices.
- You will submit the post-study questionnaire and comments and get your compensation.

If you wish to participate in the study, please:

- Sign the consent form, and
- Email it to the principal investigator balajid@dal.ca.

The principal investigator will keep the copy of this consent form.

Confidentiality Your name and address will not be required when answering any of the questions. Any of your contact information will be discarded right after the study. All your non-sensitive anonymized data will be treated confidentially and stored in a secure location. Non-sensitive anonymized data gathered from this study may potentially be used in the publications. Those data will be retained for five years after publication and then destroyed. Your personal documents will not be stored in the system

Possible follow-ups There will be no follow-ups after the end of the study. No Personal Identification will be retained and will be completely discarded after the end of the study. In all cases, all your data will be treated anonymously in all publications. You will use and test a new system which is of private intellectual property and you agree to not disclose any details about the systems that you test.

Questions Please feel free to ask the principal investigator (Balaji Dhakshinamoorthy balajid@dal.ca) about anything to do with the study before (or after) you give consent to participate.

By signing and emailing this form, you are agreeing to the following statements:

- I understand that I may discontinue my participation at any point during the study and withdrawal cannot be accepted after the end of the study.
- I have read and understood the procedure and nature of this study.
- I have had a chance to ask any questions I have about the study, and they have been answered.
- I am aware that all research material will be kept confidential
- I agree to not disclose any detail about the tested systems to any other entity.
- My quotations will be used for the evaluation and improvement of the user interface and will not be quoted in any presentations/publications.

- I understand that if I have any complaints about the experiment that I may contact

Office of Research Ethics, Dalhousie University. Email:catherine.connorsdal.ca Phone: (902) 494-1462

- I hereby consent to voluntarily take part in the study

Name: ——————

Signature: ——————

Date: ——————

Study Results

I want / do not want to receive the study results.

Email: ——————

Appendix C

Questionnaire

1. In the dissemination area (as directed in the map), which indicator makes a maximum impact for the generated index?

- Terrorism-Instru
- Import Cannabis
- Explosives
- Product Heroin
- Import OTH Drug

2. In the dissemination area (as directed in the map), which indicator makes a minimum impact for the index?

- Traffic OTH CDSA
- Human Traffic
- Import OTH Drug
- Traffic cocaine
- B&E other

3. Based on the interpretation module for the dissemination area (as directed), which explainers tend to associate positively with generated index?

- without children in a census family
- 2006 to 2010
- Natural and applied sciences and related occupations

- 20000 to 29999
- 25 to 29 years

4. Based on the interpretation result for the dissemination area (as directed), which explainers tend to associate negatively with generated Index?

- 2011 to 2016
- 25 to 29 years
- 5 or more persons
- 2006 to 2010
- without children in a census family

5. Identify the explainer that is more positively associated with the dissemination area that has higher index?

- without children in a census family
- Natural and applied sciences and related occupations
- 2006 to 2010
- Married
- 25 to 29 years

6. Identify the explainer that is more positively associated with the dissemination area that has lower index?

- 25 to 29 years
- 20000 to 29999
- 2011 to 2016
- without children in a census family
- Never married

7. In the index Comparison module, which indicator impacts more for the selection Index 1?

- Explosives
- Terrorism Instru
- B&E Firearms
- Traffic cocaine
- B&E other

8. In the index Comparison module, which indicator impacts more for the selection Index 2?

- Terrorism Instru
- Human Traffic
- B&E other
- Traffic cocaine
- Product heroine

9. In the index comparison module, identify one dissemination area where index 1 selection has more impact than index 2 selection?

- 12090588
- 12090590
- 12090623
- 12090733
- 12090871

10. In the dissemination area (as directed) Which indicator proportion has a maximum difference between index B than index A?

- B&E other
- Traffic cocaine
- Traffic OTH CDSA
- Terrorism Instru
- Arson

11. In the Dissemination area (as directed) how the addition of indicator impacts the index?

- Positively
- Unchanged
- Negatively

12. In the Dissemination area (as directed), how the removal of indicator impacts the index?

- Positively
- Unchanged
- Negatively

13. In the index comparison module, I find Index 1 has lower values than Index 2

- Agree
- Neutral
- Disagree

14. In the index comparison module, I find Index 1 has higher values than Index 2.

- Agree
- Neutral
- Disagree

Bibliography

[1] Canadian centre for justice statistics. https://www.cacp.ca/canadian-centre-for-justice-statistics.html.

[2] Google maps documentation. https://developers.google.com/maps/documentation/javascript/datalayer.

[3] Yael Albo, Joel Lanir, Peter Bak, and Sheizaf Rafaeli. Off the radar: Comparative evaluation of radial visualization solutions for composite indicators. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):569–578, 2016.

[4] Thiago Alexandre Das Neves Almeida, Luís Cruz, Eduardo Barata, and Isabel-María García-Sánchez. Economic growth and environmental impacts: An analysis based on a composite index of environmental damage. *Ecological Indicators*, 76:119–130, 2017.

[5] Kjell Andersson, Per Angelstam, Robert Axelsson, Marine Elbakidze, and Johan Törnblom. Connecting municipal and regional level planning: Analysis and visualization of sustainability indicators in bergslagen, sweden. *European Planning Studies*, 21(8):1210–1234, 2013.

[6] G.m. Antony and K. Visweswara Rao. A composite index to explain variations in poverty, health, nutritional status and standard of living: Use of multivariate statistical methods. *Public Health*, 121(8):578–587, 2007.

[7] Adithya Balaji and Alexander Allen. Benchmarking automatic machine learning frameworks. *arXiv preprint arXiv:1808.06492*, 2018.

[8] Rita Borgo, Johannes Kehrer, David Chung, Eamonn Maguire, Robert Laramee, Helwig Hauser, Matthew Ward, and Min Chen. Glyph-based visualization: Foundations, design guidelines, techniques and applications. 05 2013.

[9] Statistics Canada. Census profile, 2016 census halifax [census metropolitan area], nova scotia and nova scotia [province]. https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CMACA&Code1=205&Geo2=PR&Code2=12&SearchText=Halifax&SearchType=Begins&SearchPR=01&B1=All&GeoLevel=PR&GeoCode=205&TABID=1&type=0, Aug 2019.

[10] Statistics Canada. Uniform crime reporting survey (ucr). https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3302, Jul 2019.

[11] Chris Chatfield. *Time-series forecasting*. Chapman and Hall/CRC, 2000.

[12] Zaixu Cui and Gaolang Gong. The effect of machine learning regression algorithms and sample size on individualized behavioral prediction with functional connectivity features. *Neuroimage*, 178:622–637, 2018.

[13] Jane Elith. Machine learning, random forests, and boosted regression trees. *Quantitative Analyses in Wildlife Science*, page 281, 2019.

[14] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in neural information processing systems*, pages 2962–2970, 2015.

[15] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[16] Isabelle Guyon, Kristin Bennett, Gavin Cawley, Hugo Jair Escalante, Sergio Escalera, Tin Kam Ho, Núria Macia, Bisakha Ray, Mehreen Saeed, Alexander Statnikov, et al. Design of the 2015 chalearn automl challenge. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2015.

[17] Isabelle Guyon, Imad Chaabane, Hugo Jair Escalante, Sergio Escalera, Damir Jajetic, James Robert Lloyd, Núria Macià, Bisakha Ray, Lukasz Romaszko, Michèle Sebag, et al. A brief review of the chalearn automl challenge: any-time any-dataset learning without human intervention. In *Workshop on Automatic Machine Learning*, pages 21–30, 2016.

[18] Maike Hamann, Reinette Biggs, and Belinda Reyers. Mapping social–ecological systems: Identifying 'green-loop' and 'red-loop' dynamics based on characteristic bundles of ecosystem service use. *Global Environmental Change*, 34:218–226, 2015.

[19] Maike Hamann, Reinette Biggs, and Belinda Reyers. An exploration of human well-being bundles as identifiers of ecosystem service use patterns. *Plos One*, 11(10), Mar 2016.

[20] Ángel F. Herrera-Ulloa, Anthony T. Charles, Salvador E. Lluch-Cota, Hermán Ramirez-Aguirre, Sergio Hernández-Váquez, and Alfredo Ortega-Rubio. A regional-scale sustainable development index: the case of baja california sur, mexico. *International Journal of Sustainable Development & World Ecology*, 10(4):353–360, 2003.

[21] Patrick Hoffman, Georges Grinstein, Kenneth Marx, Ivo Grosse, and Eugene Stanley. Dna visual and analytic data mining. In *Proceedings. Visualization'97 (Cat. No. 97CB36155)*, pages 437–441. IEEE, 1997.

[22] Rob J Hyndman and George Athanasopoulos. *Forecasting: principles and practice*. OTexts, 2018.

[23] Alfred Inselberg. The plane with parallel coordinates. *The visual computer*, 1(2):69–91, 1985.

[24] Ian Jolliffe. *Principal component analysis*. Springer, 2011.

[25] David G Kleinbaum, Lawrence L Kupper, Keith E Muller, and Azhar Nizam. *Applied regression analysis and other multivariable methods*, volume 601. Duxbury Press Belmont, CA, 1988.

[26] Ilse Kotzee and Belinda Reyers. Piloting a social-ecological index for measuring flood resilience: A composite index approach. *Ecological Indicators*, 60:45–53, 2016.

[27] Klaus Krippendorff. Reliability. *The International Encyclopedia of Communication*, May 2008.

[28] Tarald O Kvålseth. Cautionary note about r 2. *The American Statistician*, 39(4):279–285, 1985.

[29] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[30] Zachary C Lipton. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

[31] Ana Carolina Lorena and André CPLF de Carvalho. Uma introdução às support vector machines. *Revista de Informática Teórica e Aplicada*, 14(2):43–67, 2007.

[32] D. F. Meyer, J. De Jongh, and N. Meyer. The formulation of a composite regional development index. *International Journal of Business and Management Studies*, 8:100 – 116, 2016.

[33] Michela Nardo, Michaela Saisana, Andrea Saltelli, Stefano Tarantola, Anders Hoffman, and Enrico Giovannini. Handbook on constructing composite indicators. *OECD Statistics Working Papers*, 2005.

[34] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[35] Randal S Olson, Nathan Bartley, Ryan J Urbanowicz, and Jason H Moore. Evaluation of a tree-based pipeline optimization tool for automating data science. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016*, pages 485–492. ACM, 2016.

[36] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, 2016.

[37] Patrick Riehmann, Manfred Hanfler, and Bernd Froehlich. Interactive sankey diagrams. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 233–240. IEEE, 2005.

[38] Jože Rovan, Kaja Malešič, and Lea Bregar. Well-being of the municipalities in slovenia. *Geodetski vestnik*, 53(1):92–113, 2009.

[39] S Rasoul Safavian and David Landgrebe. A survey of decision tree classifier methodology. *IEEE transactions on systems, man, and cybernetics*, 21(3):660–674, 1991.

[40] Josée Savoie. Neighbourhood characteristics and the distribution of crime: Edmonton, halifax and thunder bay. `https://www150.statcan.gc.ca/n1/pub/85-561-m/85-561-m2008010-eng.pdf`, 2008.

[41] Barbara Scaglia and Fabrizio Adani. An index for quantifying the aerobic reactivity of municipal solid wastes and derived waste products. *Science of The Total Environment*, 394(1):183–191, 2008.

[42] Antonio Scipioni, Anna Mazzi, Marco Mason, and Alessandro Manzardo. The dashboard of sustainability to measure the local urban sustainable development: The case study of padua municipality. *Ecological Indicators*, 9(2):364–380, 2009.

[43] Alex J Smola and Bernhard Schölkopf. A tutorial on support vector regression. *Statistics and computing*, 14(3):199–222, 2004.

[44] João Oliveira Soares, Maria Manuela Lourenço Marquês, and Carlos Manuel Ferreira Monteiro. A multivariate methodology to uncover regional disparities: A contribution to improve european union and governmental decisions. *European Journal of Operational Research*, 145(1):121–135, 2003.

[45] Lotten Wiréhn, Tomasz Opach, and Tina-Simone Neset. Assessing agricultural vulnerability to climate change in the nordic countries – an interactive geovisualization approach. *Journal of Environmental Planning and Management*, 60(1):115–134, Dec 2016.