Exploring clustering applications in outlier detection for administrative data sources

Elizabeth Ayres
Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, ON K1A 0T6

Abstract
National statistical agencies are relying more heavily on administrative data sources, which are becoming increasingly larger, requiring efficient edit and imputation procedures. Outlier detection methods currently available at Statistics Canada are highly effective in settings where the variable of interest follows a unimodal distribution, either on its own, or within groups formed by a set of class variables. Often with large administrative data sources, finding a set of class variables which can be used to satisfy this assumption is a challenge, and the effectiveness of the outlier detection is subsequently reduced. This is the case for our motivating application involving international merchandise trade data. This paper explores unsupervised clustering techniques capable of handling a mixture of quantitative and qualitative variables, with the goal of applying these techniques in order to increase outlier detection efficacy. We propose a method for using cluster analysis to isolate modal distributions as a pre-treatment to outlier detection. In addition, we examine a clustering method for outlier detection directly. These methods are contrasted with a standard approach commonly used for business surveys at Statistics Canada.

Key Words: clustering, outlier detection, feature selection, big data, machine learning

1. Introduction

The Canadian International Merchandise Trade Program (CIMTP) at Statistics Canada obtains administrative data pertaining to merchandise imports and exports from a number of sources, including the Canadian Border Services Agency, Postal Imports Control System and the National Energy Board. Collectively, this results in approximately 12 million new records with over 40 variables to process each month. CIMTP trade data is a census of international trade transactions and includes variables such as total value, quantity shipped, details about the Canadian importer, and commodity classification.

The trade data commodity classification follows the Harmonized Commodity Description and Coding System (HS). HS is an internationally standardized system where commodities are classified using a six digit nomenclature. In this system, the most general classification is at the two-digit level, and each additional pair of digits represent a more disaggregated classification. In Canada, this classification is extended to 10 digits for imports.

During processing, the trade data undergoes a number of validation as well as edit and imputation steps, prior to estimation and publishing. One of these steps pertains to the detection of outliers in the quantity variable for import transactions, where outliers are observations which are considered to fall far outside the general pattern in the data. The current process uses the relationship between value and quantity in terms of the unit value (UV) to determine outliers in quantity, where UV = value/quantity. Due to the risk of penalty for incorrect reporting, it is assumed that importers are likely to record a correct value, and therefore, an outlier in UV implies an outlier in quantity. The data is aggregated by the 10 digit HS classification, Canadian business number of the importer, and country of origin. These aggregates were determined by consultation with subject matter experts, and thought to define groups which are more homogenous with respect to UV. Within these aggregates, extreme UVs are flagged as outliers. Once outliers in UV are detected, observations undergo further processing with respect to the quantity variable. The current methodology tends to result in a considerable amount of manual review.

In order to update the current process, one approach is to utilize the same subject matter aggregates, and determine outliers using the Hidirolgou and Berthelot (1986) method. The HB method is a univariate ratio edit procedure, and is used in many business surveys at Statistics Canada. However, aggregating the data using the subject matter chosen variables does not always work well to explain trends in UV. Within these aggregates, UV often exhibits a multimodal distribution. The HB method works well for ratio variables that follow a unimodal distribution, but applying this method to variables with multimodal distributions can have undesirable effects on the results. Additionally, to apply the HB method, a set of parameters must first be specified by the user which help to determine the upper and lower bounds. This requires some investigation by the user, and it may be difficult to determine one set of parameters which will work well for the entire dataset.

Exploring ways of utilizing the inherent structure of the data to help inform outlier detection would be helpful. Cluster analysis is a classification method, where the observations are organized into subsets not known a priori, and are determined to be more closely "related" to one another than they are to observations located in other subsets based on a set of variables (Hastie et al., 2009; Xu and Wunsch II, 2009). Cluster analysis can be used for exploring data and uncovering hidden structures, and may be useful in outlier detection. One approach is to cluster observations based on a set of class variables, with the intention of creating subgroups which should be more homogenous with respect to UV. This would be followed by applying outlier detection within the clusters. Alternatively, we could cluster observations based on a set of variables, including UV, where outlying observations may fall into separate clusters on their own.

The objective of this paper is to identify clustering methods which can be applied in outlier detection for administrative data, and which address the issues faced when applying standard outlier detection methods. To do this, we begin by exploring different clustering techniques, as well as pre-clustering processes, in order to determine which methods best fit our application. Following this, we illustrate the chosen methods through a case study using the CIMTP trade data. We finish with a discussion of the results and future work.

2. Utilizing Clustering Methods with Mixed Data

Prior to performing cluster analysis on a dataset, it is typical to first carry out variable selection. Variable selection, also called feature selection, is the process of selecting a subset of most relevant variables, or features, which will be used to uncover the structure of the data. After variable selection, a proximity matrix is calculated. This matrix contains the pairwise similarity or distance measurements between each pair of observations. Last, a clustering algorithm is applied to the proximity matrix. This algorithm should be chosen carefully based on the characteristics of the clusters it will be uncovering. We discuss each of these processes, and select methods most applicable to the CIMTP trade data. To remain consistent with literature on clustering, we will use the term “features” in place of “variables” for the remainder of the paper.

2.1 Feature Selection

One of the challenges with performing cluster analysis on administrative data is managing computational complexity. As the number of observations and features increase, so does the time required for processing the data. In addition, not all of the features present may be informative for our purposes, and could result in noise. Xu and Wunsch II (2009) claim that pre and post clustering processes, including feature selection, are just as important as the clustering method itself. Therefore, it would be helpful to determine a method for selecting a subset of most informative features from the data.

Finding an appropriate method of feature selection presents another challenge; feature selection for unsupervised methods has been historically less researched than for supervised learning methods (Dash & Liu, 2000; Jain et al., 1999; Law et al., 2004). Many texts have been written about clustering that identify the importance of performing feature selection, however they give little direction on how to approach it (Anderberg, 1973; Everitt et al., 2001; Jain et al., 1999; Xu & Wunsch II, 2009).

Similar to model selection for a regression problem, we are looking to determine a subset of features which are the most informative for predicting trends in UV. With the trade data, possible predictors are class attributes with multiple degrees of freedom, and we cannot assume independence. One possible approach is to use a model selection technique for linear regression to select our subset of features. However, classical subset selection methods, such as best subset or stepwise selection are not ideal in our case. Best subset is computationally intensive for data where the number of possible predictors, p, is large. A large p can also lead to overfitting (James et al., 2017). Stepwise selection methods like backwards elimination offer an alternative which explore a restricted set of solutions, but can result in a suboptimal model. Alternatively, the LASSO method (Tibshirani, 1996) will select a model with a subset of p, but can also handle highly correlated features (Hastie et al., 2009), as seen in the trade data. It also does not require an assumption of independence. However, in a regression problem, class features would be expressed as a series of indicators, where any combination of these may or may not enter the final model depending on the method of selection. It is not informative to have only some levels of a feature included in our model.

The group LASSO (Yuan & Lin, 2006) is an extension of LASSO specifically designed for model selection using features which have multiple degrees of freedom (SAS 2017). This method, like LASSO, uses a constrained least squares problem to shrink coefficients towards a minimum of zero. Coefficients of zero effectively reduce the number of predictors. Unlike LASSO, the group LASSO method forces all levels of a feature to be either excluded from, or included in, the final model. This is desirable for our application. However, it is worth noting that this method tends to penalize features with many degrees of freedom, and therefore they are less likely to be selected. Here, we utilize the group LASSO for feature selection.

2.2 Proximity Measure

Following feature selection, an appropriate method for calculating proximity between observations must be determined. Proximity can be defined as a function of similarity or dissimilarity, and is a measurement of the pairwise distance between each observation based on a set of features (Xu & Wunsch II, 2009). The CIMTP trade dataset contains a mixture of quantitative and qualitative features, which makes determining a method for calculating distance more challenging. Jain et al. (1999) emphasizes the importance of choosing a measure carefully, which will take into account all features of differing type and scale. In order to adequately measure distance for mixed datasets, it has been suggested that a heterogeneous (Wilson & Martinez, 1997) and non-invariant (Xu & Wunsch II, 2009) calculation is best.

The Gower coefficient of similarity by Gower & Legendre (1986) will calculate the similarity between pairs of observations for any type of feature. The Gower similarity, s_1, between observations x and y is specified by

$$s_1(x,y) = \frac{\sum_{i=1}^{p} w_i \delta_{x,y}^{i} d_{x,y}^{i}}{\sum_{i=1}^{p} w_i \delta_{x,y}^{i}} \tag{1}$$

where p is the number of features, w_i is the weight of the i^{th} feature, $\delta_{x,y}^{i}$ and $d_{x,y}^{i}$ are determined based on the type of feature. For asymmetric categorical features, where x_i and y_i are the values of the i^{th} feature for observations x and y, respectively

$$\delta_{x,y}^{i} = \begin{cases} 1, & if\ either\ x_i\ or\ y_i\ are\ present \\ 0, & if\ both\ x_i\ and\ y_i\ are\ absent \end{cases}$$

and

$$d_{x,y}^{i} = \begin{cases} 1, & if\ x_i = y_i \\ 0, & if\ x_i \neq y_i' \end{cases}$$

while $\delta_{x,y}^{i} = 1$ and $d_{x,y}^{i} = 1 - |x_i - y_i|$ for ratio types. Asymmetric categorical refers to a categorical feature where $x_i = y_i$ is more informative than $x_i \neq y_i$. Using our data as an example, two observations having the same HS10 classification is more informative to us then if they do not share the same classification. Note that the Gower method also calculates $\delta_{x,y}^{i}$ and $d_{x,y}^{i}$ for symmetric categorical and ordinal features, but these are not applied in our case study. One of the benefits of the Gower similarity measure is that it allows for individual weighting of characteristics, specified by w_i, which could prove useful for further optimization of cluster structures. Lastly, this measure does not require recoding of quantitative features to a series of indicators (Gower & Legendre, 1986), making it more efficient for large datasets.

2.3 Cluster Analysis

The underlying assumptions of a particular clustering algorithm will dictate which cluster configurations the analysis is able to detect with regards to size, shape and dispersion (Jain et al., 1999; Xu & Wunsch II, 2009). If not chosen carefully, clustering could impose a structure on the data, rather than reveal an existing one (Everitt et al., 2001).

For example, K-means is a well-known clustering method which assigns each observation to one of *K* clusters based on minimizing a squared error criterion, where *K* is specified by the user a priori. One of the major drawbacks of this method is that K-means relies on a random initial cluster assignment. Different configurations could lead to different optimal solutions and results in greater variability for this method (Jain et al., 1999¸ Xu & Wunsch

II, 2009). Moreover, how to select *K* is not clear, and different choices of *K* could drastically alter the results (Everitt et al., 2001; Xu & Wunsch II, 2009).

For the CIMTP trade data, where UV tends to exhibit a skewed multimodal distribution, choosing a method which makes assumptions about the distribution of the data would not be an ideal approach. Furthermore, due to the size and complexity of the data, parameter selection would need to be flexible and automated. Density based clustering methods, such as mode analysis and nearest neighbor procedures, have the advantage of scalability, and are capable of detecting clusters of unequal size which are elongated, or irregularly shaped (Xu & Wunsch II, 2009). A density based approach could be useful for our data where the number of clusters, as well as cluster size and shape, would be hard to determine in advance.

Another technique which could be helpful is hierarchical clustering. Agglomerative hierarchical clustering is analogous to an inverted classification tree method. With agglomerative hierarchical clustering, all observations begin in a cluster on their own (ie. the leaves of the tree), and based on some clustering criterion, clusters merge together one at a time forming the branches of the tree. The process terminates once all clusters have merged, forming one cluster containing all observations (ie. the trunk of the tree). Due to the way the hierarchical cluster structure is built this method tends to be sensitive to outliers, particularly when using the single linkage clustering criterion (Xu & Wunsch II, 2009). For single linkage, the similarity between two clusters is equivalent to the similarity between the most similar members of each cluster. As a result, single observations that merge closer to the trunk of the tree tend to be vastly different from the other observations. This property could be useful in performing outlier detection.

3. Application of Outlier Detection Methods to CIMTP Data

To illustrate an application of cluster analysis in outlier detection using the techniques we describe above, we present two methods. These methods are applied to the CIMTP trade data, along with a standard method of outlier detection for comparison purposes.

3.1 Hidiroglou-Berthelot with Subject Matter Groups

This approach is commonly applied to business surveys at Statistics Canada, and will be referred to as Method 1. The data is first grouped by the 10 digit HS code, the Canadian business number of the importer, and the country of origin. Then the HB method is applied to each group in order to determine outliers in UV. Observations with an outlying UV are considered to have a correct value, and therefore an outlying quantity.

The HB method calculates acceptance boundaries based on the quartiles of the distribution for the feature of interest, and flags observations which fall outside of these thresholds. The focus of this paper is on the application of clustering methods to outlier detection, and so we will not expand further on the HB method. For further details on this method, see Hidirolgou and Berthelot (1986).

3.2 Hidiroglou-Berthelot with Non-Parametric Density Clustering

This approach, referred to as Method 2, utilizes a non-parametric density clustering algorithm available in the SAS function PROC MODECLUS. This algorithm is used to assign observations into clusters, then the HB method is applied to each cluster in order to detect outliers in UV. Since the quantity variable is published by HS6, country of origin and province of clearance, we first subset the data by HS6. Feature selection is applied separately for each HS6, where country of origin and province of clearance are included in the list of possible predictors. Feature selection is performed using the group LASSO method. In order satisfy the normality assumption, a log, inverse, and square root transformation are applied separately to UV, and each transformation is visually assessed by histogram. Generally, it is observed that the $\log(UV)$ transformation best meets this assumption for CIMTP data.

Feature selection is performed using PROC GLMSELECT with 50 iterations and five-fold external cross validation. The constrained least squares problem for the group LASSO is solved using the Lagrangian form

$$min\|\boldsymbol{y} - \boldsymbol{X\beta}\|^2 + \lambda \textstyle\sum_{i=1}^{p} \sqrt{|G_i|}\left\|\beta_{G_i}\right\| \tag{2}$$

where $\boldsymbol{y}$ denotes the response, $\boldsymbol{X}$ is the matrix of covariates, $\boldsymbol{\beta}$ is the vector of coefficients which will be minimized, $\|\cdot\|$ denotes the Euclidean norm, p is the number of features, $\sqrt{|G_i|}$ is the group weight, and is equal to the square root of the number of group coefficients, β_{G_i}, which correspond with group i, and λ is the Lagrangian multiplier. SAS uses a method by Nesterov (2013), which proposes that a sequence of regularization parameters $\rho, \rho^2, \ldots$ be used to determine λ, where $0 < \rho < 1$.

Following feature selection, a square $n \times n$ proximity matrix is calculated using the DGOWER distance measure in PROC DISTANCE. The DGOWER distance measure is defined as

$$d_2(x, y) = 1 - s_1(x, y), \tag{3}$$

where $s_1(x, y)$ is the Gower similarity (eq. 1) between observations x and y. Features are standardized and given a weight, w_i, of 1 (see eq. 1).

Clustering is carried out using a non-parametric density algorithm in PROC MODECLUS. This procedure first calculates a non-parametric density estimate $\hat{f}$ for the m^{th} observation, x_m, using the following formula

$$\hat{f}_m = \frac{n_m}{n v_m}, \tag{4}$$

where n is the total number of observations, v_m is the volume of the neighbourhood x_m which contains its k nearest neighbors, and $n_m = k + 1,$. Cluster membership is then assigned based on iteratively merging observations, or clusters of observations, to the nearest neighboring cluster with a greater density. k is specified using the following formula, $k = \lceil 0.1 \times \mathrm{n}^{0.8} \rceil$, which is based on a recommendation in the SAS (2017) documentation. Lastly, the HB method is applied to each cluster to determine outliers in UV.

3.3 Agglomerative Hierarchical Clustering

The last approach, referred to as Method 3, applies clustering alone as a method of outlier detection, and uses the same procedure of feature selection that we use for Method 2. However the distance calculation differs slightly under this approach.

Similar to Method 2, a lower triangular $n \times n$ proximity matrix is calculated using eq. 3, based on all features selected in the model. Importantly, for Method 3 we include UV in this calculation. As we are interested in detecting outliers in UV to inform outlier detection in quantity, UV is weighted more heavily. Further work on the impact of the weights is required.

Clustering is implemented using an agglomerative hierarchical clustering algorithm in SAS PROC CLUSTER. The density procedure uses a non-parametric probability estimation based on a user specified number of nearest neighbors, k. As in Method 2, we use $k = \lceil 0.1 \times n^{0.8} \rceil$. The resulting output dataset from PROC CLUSTER contains details of the entire tree structure, which will be used to select the outlying clusters. The following steps are based on an outlier detection method proposed in Loueiro et al. (2004). PROC TREE is used to dissect the tree at a specified number of clusters, nc, where $nc = \max(2, \lceil 0.2n \rceil)$. If any of these nc clusters contain only 1 observation, then it is considered an outlying cluster.

4. Results

We now apply Methods 1, 2 and 3 to a subset of the CIMTP import trade data. This subset consists of data for one month with an HS4 classification of 8511, which includes trade transactions related to ignition and starting equipment. This subset contains 19,754 records from HS6 classifications 851130, 851140, and 851150. We choose this subset as it is the most recent dataset available to us. The intention of this case study is to characterize the resulting sets of outliers so that we may gain insight into how well these methods are performing.

4.1 Results of Feature Selection

First, we examine the results of the feature selection. For Method 1, all data is aggregated by the same three features: HS10, business number, and country of origin. For Methods 2 and 3, data is first subset by HS6, then feature selection is carried out independently within each subset.

Table 1 displays a comparison of the subject matter chosen features we use in Method 1, and the results of the feature selection we use in Methods 2 and 3 for each of the HS6 subsets. For Methods 2 and 3, we observe some overlap between the features chosen by the group LASSO method for each subset. Despite this overlap, we also observe that different models were chosen overall. This latter observation indicates that a more flexible approach to feature selection may be necessary for our data, especially considering that feature selection was applied only to one specific commodity type.

In comparing the features used in Method 1, only HS10 was selected in any of the models for Methods 2 and 3. There are a couple of possible explanations for this. First, it could be that business number and country of origin are not good predictors of UV compared to the other features present in the data. Another possibility is that since country of origin and business number both have a relatively large number of levels, it is possible that these features were left out of the model due to the penalty on the least squares constraint discussed in Section 2.1.

Table 1: A Comparison of Subject Matter Chosen Features used in Method 1, and Features Selected by Group LASSO used in Methods 2 and 3.

Method 1	**Methods 2 and 3**		
All Data	851130	851140	851150
HS10	Customs Office Region	Customs Office Region	Customs Office Region
Business Number	Value for Duty Code	Value for Duty Code	Value for Duty Code
Country of Origin	Entry Type	Entry Type	Entry Type
	Sales Rate	Sales Rate	Sales Rate
	HS10	Mode of Transport	Region of Export
			Tariff Code

4.2 Outlier Detection

Next, we examine the outlier detection results. First, we compare Methods 1 and 2 to demonstrate the effect of using clustering in place of the subject matter chosen groups. Table 2 is a two-way frequency table of outliers versus non-outliers in the quantity variable by method. We observe that in changing the approach to aggregating the data prior to applying the HB method, Method 2 flags over three times the number of outliers as Method 1. The majority of outliers are detected by one method and not the other, with Method 1 flagging 96 outliers uniquely, and Method 2 flagging 450. We are interested in characterizing these unique subsets.

Table 2: A Comparison of Outlier and Non-Outlier Counts by Methods 1 and 2.

		Method 2		
		Non-Outlier	*Outlier*	*Total*
Method 1	*Non-Outlier*	19150	**450**	19600
	Outlier	**96**	58	154
	Total	19246	508	19754

One way to characterize these subsets is to consider how influential these outliers are relative to non-outliers, based on the domains of interest. Influential observations are not necessarily outliers, however, outliers which are *also* influential are of interest to us, and should be investigated further. Quantity estimates are published by HS6, country of origin and province of clearance. To determine influence, the relative mean absolute difference (RMAD) is calculated for each observation by domain. RMAD for the m^{th} observation, x_m, is calculated using the following equation:

$$RMAD_m = \frac{|\bar{x}_\ell - \bar{x}_\ell{}^*|}{\bar{x}_\ell},$$

where $\bar{X}_\ell$ is the mean quantity for domain of interest ℓ, and $\bar{X}_\ell^{\,*}$ is the mean quantity for domain of interest ℓ excluding observation x_m.

Table 3: The Distribution of Outliers Detected Uniquely with Methods 1 and 2 by Rank of Influence on the Domain.

		Rank of Influence on Domain			
	Total Unique Outliers	*Top 1%*	*>1% to ≤ 5%*	*> 5% to ≤ 10%*	*Bottom 90%*
Method 1	**96** 100.00%	6 6.25%	3 3.13%	5 5.21%	82 85.42%
Method 2	**450** 100.00%	14 3.11%	25 5.56%	39 8.67%	372 82.67%

Table 3 displays the frequency and percentage of outliers which are detected uniquely by each method, and which rank in the top 1%, > 1% to ⩽ 5%, and > 5% to ⩽ 10% most influential observations. Although Method 2 detects more of the top influential observations by count, looking at the percentages we observe that the distribution is quite similar to those selected uniquely by Method 1. Changing the manner in which data is aggregated does not appear to result in the selection of more influential or less influential outliers.

Next, we compare Methods 2 and 3, which both use the same feature selection procedure, but differ in how outlier detection is performed. Table 4 is a two-way frequency table of outliers versus non-outliers in the quantity variable by method. Table 4 shows that Methods 2 and 3 detect a similar number of outliers, with Method 2 detecting more outliers than Method 3. Similar to table 2, the majority of outliers are identified uniquely, with Method 2 detecting 379 uniquely, and Method 3 detecting 265.

Table 4: A Comparison of Outlier and Non-Outlier Counts by Methods 2 and 3.

			Method 3	
		Non-Outlier	*Outlier*	*Total*
	Non-Outlier	18981	**265**	19246
Method 2	*Outlier*	**379**	129	508
	Total	19360	394	19754

Table 5 displays the frequency and percentage of outliers which are detected uniquely by each method, and which rank in the top 1%, > 1% to ⩽ 5%, and > 5% to ⩽ 10% most influential observations. By count, both methods appear to detect similar numbers of influential units. However, more of the outliers which are identified by Method 3 are influential, where roughly 35% of these fell within the top 10%, compared to roughly 21% by Method 2. This is in contrast to what is observed in table 3, where the distribution of unique outliers was similar between methods. Based on the distribution of outliers uniquely detected by each method over rank of influence on the domain of interest, Method 3 appears to perform best overall.

Table 5: The Distribution of Outliers Detected Uniquely with Methods 2 and 2 by Rank of Influence on the Domain.

		Rank of Influence on Domain			
	Total Unique Outliers	*Top 1%*	*>1% - ≤ 5%*	*>5% - ≤ 10%*	*Bottom 90%*
Method 2	**379** 100.00%	14 3.69%	25 6.60%	39 10.29%	301 79.42%
Method 3	**265** 100.00%	12 4.53%	41 15.47%	41 15.47%	171 64.53%

5. Discussion

In this paper, we examine the application of clustering in outlier detection for administrative data, and illustrate this application through a case study involving the CIMTP trade data. We discuss the challenge of using a set of features based on subject matter knowledge, which attempt to explain trends in UV for the full set of data. The methods we propose in this paper do not rely on subjective selection of features. As well, feature selection can be performed separately for different subsets of data, and is therefore a more flexible approach than the current method. Feature selection can also be automated, which is ideal for large administrative datasets. We also identify the challenges involved with parameter selection for the HB method, and that it may be difficult to select a set of parameters which works best for all of the data. Cluster analysis is an unsupervised method which uses the inherent structure of the data to inform the way the data is grouped, and is therefore more objective than standard methods. Importantly, parameter selection for clustering could be automated and data driven. Lastly, Method 3 could be adapted to perform multivariate outlier detection, where the HB method is strictly univariate.

In this paper, we also discuss the many considerations that must be made prior to applying clustering. Based on these considerations, we attempt to select techniques which are best suited for an application in outlier detection on CIMTP trade data. We use a case study to examine these techniques, and reveal Method 3 as the best performer based on the results of our investigation. However, we do understand that this paper does not present the entire picture; more work is required in characterising the subsets of outliers detected uniquely by each method. As we are working with real data, the only way to verify the accuracy of the results would be through a thorough manual review. Therefore, it is essential that we look for ways to evaluate methods which do not rely on such extensive processes.

The outlier detection methods we propose in this paper involve specifying several parameters. Each of these parameters can potentially alter the final results, and therefore parameter selection must be informed by the structure of the data and context of the problem. The parameter values we select here are based largely on suggestions in the literature, with minor adjustments to better fit the CIMTP trade data. However, few studies have been completed on how to select the optimal parameters. Therefore, an important area of future research is the automation of parameter selection, for example, calculating the number of nearest neighbors, k, as a function of the number of observations.

With respect to the techniques we apply in this paper, we identify one potential drawback. The group LASSO method may not be ideal for our data, as it tends to exclude features with higher degrees of freedom, which may impede outlier detection if these features are in fact important. Other methods of feature selection should be considered.

Lastly, the DGOWER distance measurement allows for the weighting of features, which should be further explored. Feature weighting could greatly impact how outliers are detected. For instance, if features are left unweighted for the distance calculation in Method 3, outliers in other features could be identified, such as a unique business number. A separate investigation into how this method could be applied to multivariate outlier detection would be worthwhile. With respect to the clustering methods outlined in this paper, further exploration is required into the effect of handling "ties".

Acknowledgements

I would like to acknowledge the following individuals for their contributions to this project: Nada Habli, who made the initial suggestion of looking at cluster analysis, and provided much support on developing my understanding of the subject matter material; Michel Hidiroglou and Wesley Yung, who were both helpful and supportive, offering ideas and providing feedback during the development of this project; and Shaun Pinder, who was the first methodologist responsible for looking into ways of updating the quantity edit and imputation module, and wrote the initial programs for the Hidiroglou-Berthelot method.

References

Anderberg, M.R. (1973). Cluster Analysis for Applications. New York, New York, USA: Academic Press, Inc.

Dash, M., & Liu, H. (2000). Feature Selection for Clustering. In: Terano T., Liu, & H., Chen A.L.P. (2000) Knowledge discovery and data mining: current issues and new applications. 4th Pacific-Asia Conference, PAKDD 2000, Proceedings, Lecture Notes in Computer Science, vol 1805. Berlin, Heidelberg: Springer.

Everitt, B. S., Landau, S., & Leese, M. (2001). Cluster Analysis. London: Arnold.

Gower, J. C., and Legendre, P. (1986). "Metric and Euclidean Properties of Dissimilarity Coefficients." Journal of Classification 3:5–48.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.). New York, New York, USA: Springer Science+Business Media. doi:10.1007/b94608.

Hidiroglou, M.A., & Berthelot, J.-M. (1986). Statistical Editing and Imputation for Periodic Business Surveys. Survey Methodology, 12(1), 73-83.

Jain, A., Murty, M., & Flynn, P. (1999, September). Data Clustering: A Review. ACM Computing Surveys, 31(3), 264-323.

James, G., Witten, D., Hastie. T., & Tibshirani, R. (2017). An Introduction to Statistical Learning with Applications in R. New York, New York, USA: Springer Science+Business Media. doi:10.1007/978-1-4614-7138-7.

Law, M., Figueiredo, M., & Jain, A. (2004). Simultaneous feature selection and clustering using mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26(9): 1154-1166.

Loureiro, A., Torgo, L., & Soares, C. (2004). Outlier Detection Using Clustering Methods: a data cleaning application. Proceedings of KDNet Symposium on Knowledge-based Systems for the Public Sector. Bonn, Germany.

Nesterov, Y. (2013). Gradient methods for minimizing composite objective function. Mathematical Programming , 140(1), 125-161.

Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. Journal of the Royal Statistical Society, 58(1), 267-288. Retrieved January 26, 2018, from http://www.jstor.org/stable/2346178

SAS Institute Inc 2017. SAS 9.4 Help and Documentation, Cary, NC: SAS Institute Inc.

Wilson, D. R., & Martinez, T. R. (1997, January). Improved Heterogeneous Distance Functions. Journal of Artificial Intelligence Research, 6, 1-34.

Xu, R., & Wunsch II, D. C. (2009). Clustering. Hoboken, New Jersey, United States: John Wiley & Sons, Inc.

Yuan, M., & Lin, Y. (2006). Model selection and estimates in regression with grouped features. Journal of the Royal Statistical Society, 68(1), 49-67. Retrieved January 26, 2018, from http://www.jstor.org/stable/3647556