Outlier Detection Methods for mixed-type and large-scale data like Census

**Keywords:** Outlier Detection, Anomalies, Census, Mixed Data, Large Scale

# Introduction

Outlier detection (OD) refers to the problem of finding patterns in data that do not conform to expected normal behaviour. OD has been a widely researched problem and finds immense use in a wide variety of application domains. In this paper we consider the domain of building automated OD methods for quality assure the next 2021 UK Census.

The scale and nature of such dataset pose computational challenges to traditional OD methods. In general, the scale of the full Census is too large for a sequential execution of the OD methods. Most of the methods scale super-linearly with the size of the dataset and need either a distributed implementation or separate runs of the algorithm on chunks of the dataset. Additionally, Census questions are of mixed type (numeric, categorical, ordinal, free-text and date) and detecting outliers in this multi-dimensional space is an open area of research with no optimal solution yet.

Experiences from previous census processing show that it is easy to be overwhelmed with data quickly, and a mechanism for pointing in the right direction will save huge amounts of time and improve quality where it is needed most. It will also help minimize the risk of serious errors, identifying them earlier.

## Outliers in Census data

Outliers, as defined earlier, are patterns in data that do not conform to a well-defined notion of normal behaviour. For the purpose of quality assuring the next Census, we can identify the following broad categories of outliers of interest:

* Unusual individual records, i.e. a single observation (e.g. A 100 year old student with active employment)
* Groups of records with unusual characteristics (e.g. A high number of people with a manager occupation in certain geography)
* Impossible values/combinations.
* Spurious distributions for subsets of the data (e.g. A highly skewed distribution for given demographics in given area)
* Other problems that we might not know about. This could be particularly relevant for Census 2021, as for the UK this is going to be the first Census delivered online, thus raising new/different types from those arose in past paper-filled censuses.

Outliers in Census has in the past arose due various reasons, including errors in the questionnaire (e.g. misunderstood question, typos), scanning errors (e.g. incorrect recognition of the scanned text), coding errors (e.g. incorrect geography coded for the workplace), imputation errors (e.g. imputing data from erroneous records) or introduced at later steps of Census processing. Nonetheless, a survey of human population features such as Census may also include anomalies that are purely natural, although the reading is still worth flagging for verification to ensure no errors has been made.

In the remainder of this paper, we categorise and analyse different categories of OD methodologies that could be applicable given the challenges posed by the data and the environment in which the research has been carried. This work is being carried out with the goal to develop a set of lightweight tools which could be run against the full-scale census data to automatically flag anomalous observations in the dataset. This could be incorporated in the census processing pipeline at a number of different stages, and help focus resources efficiently.

# Methods

Following a thorough literature review carried out by the University of Southampton [1], we investigated some of the state the art methodologies, focusing our study on the following aspects:

* Time and space complexity: The raw Census 2021 data is expected to contain ~ 65 million records on > 50 variables of mixed type, which expanded lead to roughly 200+ variables. Selected methods need to be conscious on both the amount of time as well as space or memory taken by an algorithm to run as a function of the length and dimensionality of the input.
* Existence of a distributed version of the algorithm: Existing ready-to-use distributed implementations for OD methods are rare, although there are distributed implementations of part of algorithms (e.g. frequent itemset mining, clustering) which could be used either directly or with modifications of their source code.
* Ability to handle high dimensional data
* Ability to handle mix-typed data
* Existence of pseudocode/implementation: Preference is given in those situation where an implementation exists as these limits the amount of effort needed. Off-the-shelf Python implementations are, however, available only for the sequential versions of some of the algorithms.

Using the above criteria, the below listed potential algorithms for OD were selected for implementation.

## Statistical/Machine Learning methods

There is certainly some overlap between classical statistical methods and the data science approaches. We focus on statistical methods which do not involve learning or training. These are:

* Simple univariate Probabilistic Anomaly Detector (SPAD) [2] estimates multivariate probability as a product of univariate probabilities.
* Pattern based Outlier Detection uses Logistic Regression (POD) or one-class SVM (POD\_SVM) [3] to learn patterns and then formulate the outlier factor in mixed attribute datasets.
* Isolation Forest (iForest) [4] employs an isolation mechanism using trees to isolate every instance in the data. The intuition is that anomalies are more susceptible to isolation and therefore have shorter average path lengths than normal instances.
* MIXed data Multilevel Anomaly Detection (MIXMAD) algorithm [5] sequentially constructs an ensemble of Deep Belief Nets (DBNs) with varying depths. Similar concepts is can be replicated using autoencoders or substituted by the newest Generative Adversarial Networks (GANs) [6] based on deep neural net architectures.
* KAMILA (KAy-means for MIxed LArge data sets) [7] An iterative clustering method that uses kernel density estimation to flexibly model spherical clusters in the continuous domain and a multinomial model in the categorical domain.

## Neighbourhood based methods

Defining outliers by their distance to neighbouring examples is a popular approach to finding unusual examples in a data set. It generally requires a distance measure to compute all pairwise distances between instances. Some of the proposed methods below overcome this expensive computation by limiting the search to a restricted number of neighbours. In the mixed-type domain, Euclidean distance is generally used for numerical attributes, while Occurrence Frequency was suggested in the categorical domain.

* Local Outlier Factor (LOF) method [8] is based on relative density of an instance with respect to its *k*-neighbourhood.
* Random subsample (SP) [9] search the nearest neighbour in a small random subsample to detect anomalies
* Simple nested-loop algorithm with a simple pruning rule (SNLA\_SPR) [10] can give near linear time performance when the data is in random order and a simple pruning rule is used

## Frequent Itemsets based methods

Frequent Itemsets based methods detects outliers assuming that, in the categorical space, outliers are points with highly irregular or infrequent values. In the mixed space, an outlier would be irregular or anomalous in either categorical, continuous or both domains.

* Outlier detection for mixed attribute datasets (ODMAD) [11] computes an anomaly score for each point taking into consideration the irregularity of the categorical values, the continuous values, and the relationship between the two spaces in the dataset

# Discussion

A literature review was conducted on the major statistical and data science methods for detecting outliers and anomalies in census data. These were condensed under three categories which are:

1. Statistical/Machine Learning methods
2. Neighbourhood based methods
3. Frequent Itemsets based methods

This work is being carried out and will culminate with the development a set of lightweight tools to be ready to test on the mid-2019 UK Census rehearsal. Ultimately, these could be run against the full-scale 2021 Census data in a distributed fashion to automatically flag anomalous observations in the dataset. The up to date results of experiments with such methods will be the main focus of the presentation.

Outlier detection is an extremely important problem with direct application in a wide variety of domains. The opportunities of outlier detection are huge. Quality assurance is one obvious application. Our aim is to design not only methods working on Census data, but that can be re-used for other, especially survey-like data. Also, outlier detection is essentially a synonym for novelty detection, as it can be looked at as a technique for finding previously non-existent types of records. Finally, many of the OD methods can be useful to provide more insight into the distribution of the data (probability-based methods), or their structure (e.g. clustering based methods).

# References

1. Z. Sabeur, P. Smith, J. Dawber, G. Correndo, G. Veres and A. May, Outliers and Anomalies Detection Methods with Applications to the 2021 Census, Internal document available on request (2018).
2. S. Aryal, K.M. Ting and G. Haffari, Revisiting Attribute Independence Assumption in Probabilistic Unsupervised Anomaly Detection, PAISI (2016).
3. M.S. Mausam, R. Bart, S. Soderland and O. Etzioni, Open language learning for information extraction, Empirical Methods in Natural Language Processing and Computational Natural Language Learning (2012), p. 523-534.
4. L.F. Tony, T.K. Ming and Z. Zhi-Hua, Isolation-based anomaly detection, ACM Transactions on Knowledge Discovery from Data (TKDD) 6.1 (2012), 3.
5. K. Do, T. Tran and S. Venkatesh, Multilevel Anomaly Detection for Mixed Data (2016).
6. F. Gökgöz, Anomaly Detection using GANs in OpenSky Network (2018)
7. A. Foss, M. Markatou, B. Ray and A. Heching, A semiparametric method for clustering mixed data, Machine Learning (2016).
8. M.M. Breunig, H.P. Kriegel, R.T. Ng and J. Sander, LOF: identifying density-based local outliers, ACM SIGMOD Conference on Management of Data (2000), p. 93–104.
9. M.E. Otey, A. Ghoting and S. Parthasarathy, Fast Distributed Outlier Detection in Mixed-Attribute Data Sets, Data Mining and Knowledge Discovery (2006), 12: p. 203–228.
10. S.D. Bay and M.A. Schwabacher, Mining distance-based outliers in near linear time with randomization and a simple pruning rule, ACM SIGKDD international conference on Knowledge discovery and data mining (2003).
11. A. Koufakou and M. Georgiopoulos, A fast outlier detection strategy for distributed high-dimensional data sets with mixed attributes, Data Mining and Knowledge Discovery (2010), 20: p. 259–289.