Data Science and Engineering (2024) 9:62–72
https://doi.org/10.1007/s41019-023-00234-7

RESEARCH PAPER

Anomaly Detection with Sub-Extreme Values: Health Provider Billing

Rob Muspratt[1,2] · Musa Mammadov[1]

Received: 12 May 2023 / Revised: 23 October 2023 / Accepted: 27 October 2023 / Published online: 29 November 2023

Abstract
Anomaly detection within the context of healthcare billing requires a method or algorithm which is flexible to the practicalities and requirements of manual case review, the volumes and associated effort of which can determine whether anomalous output is ultimately actioned or not. In this paper, we apply a modified version of a previously introduced anomaly detection algorithm to address this very issue by enacting refined targeting capability based on the identification of sub-extreme anomalies. By balancing the anomaly identification process with appropriate threshold setting tailored to each sample health provider discipline, it is shown that final candidate volumes are controlled with greater accuracy and sensitivity. A comparison with standard local outlier factor (LOF) scores is included for benchmark purposes.

Keywords Healthcare provider · Anomaly detection · Outlier thresholds · Sub-extreme values

1 Introduction

The Transport Accident Commission of Victoria (TAC) is a State Government owned organisation whose key function is funding treatment and support services for people injured in transport accidents. The TAC generates health provider billing transactions as a by-product of processing and funding health provider accounts and services for its clients. It is the descriptive attributes of these transactions which constitute the scheme output variables of this study (e.g. service item selection).

Use of the term anomaly and the range of alternate descriptors available depends highly on the domain of application, e.g. outliers, discordant observations, exceptions, aberrations, surprises, peculiarities or contaminants [1]. Whilst original research upon which this investigation is based was concerned with defining outliers, it is more appropriate to apply the anomaly descriptor in this instance. The term "anomaly" more accurately reflecting the "non-extreme" nature of discordant observations sought [2, 16]. The novel application in this context is the non-parametric classification of anomalies which transgress more traditional definitions offered for outliers in the literature [12]. Adapting the notion of outliers to anomalies [13, 14] it can also be stated that whilst all extreme values are anomalies, not all anomalies are necessarily extreme values.

Consideration of appropriate output translation and targeting of health provider–client combinations along with computational requirements and reusability led to the development of a bespoke method based on direct provider comparison [6]. In terms of actions or behaviours over similar cases/claims, the method utilises the following scheme:

Input → Provider → Output

Input is assumed to combine a set of features that could be used to define "similar clients", for example, age, gender, postcode, injury types, etc. Output is assumed to be a set of responses/actions by a particular health provider, for example this may include service types, service levels, service intensities, billed amount, type of billing, etc. In this paper, we develop further the approach introduced in [6, 8] by considering and addressing the following two important issues with regard to encoded service levels:

- Unusual patterns are evident in the encoding of service levels which require decomposition and upon application of a best-fit distribution make clear intervals of divergence in the tail.

✉ Rob Muspratt
rmuspratt@deakin.edu.au

Musa Mammadov
musa.mammadov@deakin.edu.au

[1] School of Information Technology, Deakin University, Melbourne, VIC 3125, Australia

[2] Transport Accident Commission, Geelong, VIC 3220, Australia

- Accurate determination of the thresholds associated with these intervals of divergence is required to isolate the non-extreme anomalies of interest.

The persistent challenge in identifying a sub-extreme interval for the purposes of anomaly detection in this particular business context is balancing the volume of output with the degree of abnormality observed [3]. The scheme or model employed needs to be flexible to this requirement [5] and allow for tuning via the use of an appropriate threshold parameter. We will show that this is indeed possible by applying the suggested algorithm and explore the resultant intervals that are derived by said algorithm in the process. In addition, comparison with a known benchmark algorithm described in the literature [9, 10, 15] gives perspective on the complexity of the input data elements used in this study and the need for appropriate application of business knowledge to aid final interpretability.

2 Algorithm and Notations

Distance functions for input (x) and output (y) variables will be denoted by d^x and d^y, respectively.

2.1 Input/Output Spaces

Consistent with previous analysis [4, 6, 11], the following input variables are used to select similar claims among all sample data points:

I1: List of injuries is a list of all 20 possible client injuries. The resulting injury vector

$I = (I_1, \ldots, I_{20})$ is a binary vector of length 20 representing corresponding incidence of injury. Depending on the severity of injuries, subsequent weightings are applied in the form:

$I_k = 100\, I_k/k^2$, $k = 1, \ldots, 20.$

I2: Variable "Age", denoted by A, has values in the interval (0,100), these values are rescaled to the range [0, 5].

I3: Variable "Time from Accident", denoted by T, is also rescaled to the range [0, 5].

After scaling all the input variables $x = (I, A, T)$ the Euclidean distance would be the best choice to define a neighbourhood in the input space around a given data point in D, defined as the distance function d^x. Weighting and scaling has been applied to standardise each input variable such that no one input characteristic takes dominance based on its intrinsic scalar value (e.g. age value will always be greater than time from accident) apart from severe injuries reflected by $k < 4$.

The output variable considered in this paper is the following:

O1: Service levels is a vector of services $s = (s_1, \ldots, s_L)$ defined for each provider–claim combination where L is the number of service levels, and s_l is the number of services of level $l \in \{1, \ldots, L\}$.

Service levels are derived from pre-existing Medicare Benefit Schedule (MBS) [7] categorisations or TAC service item definitions based on consultation time and/or complexity as appropriate to the related health discipline. When considering service levels, measuring the proportions of billing at each level is of most interest, accordingly a Cosine measure is best in this case. Therefore, distance d^y between two services s^1 and s^2 will be defined in the form:

$$d^y(s^1, s^2) = 1 - \mathrm{Cos}(s^1, s^2)$$

2.2 Variation from Local Mean (VLM)

The initial algorithm used in this paper is described by Steps 1–4 below, where Steps 1, 3 and 4 are adopted from [6] and Step 2 from [8].

Step 1: Given a health provider–claim (or patient) combination (p, c), the degree of abnormality with regard to clinical treatment billed is calculated thus. Consider an arbitrary data point $(p^0, c^0; x^0, y^0)$ in our domain, D.

- Calculate the distance $d^x(x^0, x)$ from all data points $(p, c; x, y) \in D$ and select the closest n^{top} points, the neighbourhood, that will be denoted by N^0.
- Calculate the average value AvS^1 of distance $d^y(y^0, y)$ over all data points in N^0.
- The resultant outlying value $\mathrm{VLM}(p^0, c^0) = AvS^1$

Note: This resultant value defines anomalies in terms of the "local" neighbourhood, that is, the divergence with respect to the closest n^{top} claims; we will call it *variation from local mean* (VLM).

Step 2: In the case of service levels, decomposition of D is required as described in [8] to produce clusters of interest, namely Modal, Specialised and High (as a sub-component of Aberrant), denoted here as D_M, D_S and D_H.

Step 3 (Best Fit): Find the best-fit distribution function for the value VLM over data D_M, D_S and D_H as required.

Step 4: Define and select appropriate threshold values/intervals for service levels based on clusters of interest.

Finding the best threshold for extreme outlying values on the right tale of the distribution function is an important but difficult problem. The best fit found in **Step 3** is used in this step by analysing the divergence between the best fit and related variable (VLM). It reveals two important points:

- There is often a clear threshold point at the right tale where this divergence occurs.

- There are two intervals on the right tale after the threshold value, namely sub-extreme (SE) and high extreme (HE).

2.3 Defining Interval Sub-Extreme (SE)

The main rationale in considering the top two subsets SE and HE is as follows. Practice shows that the highest ranked anomalies usually have solid reasons for their large outlying values (as exceptional cases or data errors). Accordingly and as a result of past TAC experience, the most interesting anomalies can also be expected in the range of SE rather than in HE. This is often a reflection of providers maximising their return with regard to billing behaviour whilst stopping short of becoming true outliers amongst their peers. Application of a best-fit distribution to the variable of interest generally reveals a clear divergence that exists in the right tail and highlights the presence of sub-extreme anomalies. An interesting point is that this interval does not constitute the whole right tail (say all points above some threshold on the right tale) which is the most common approach when defining outliers in the literature. This justifies dividing the extreme right tale into two intervals SE (that is, $[t_1, t_2]$) and HE (that is, $[t_2$, infinity), and searching for anomalies in the section SE rather than in HE.

After finding the best fit in **Step 3**, values t_1 and t_2 **c**ould be defined as a solution to some optimisation problem described in the form:

Optimise (w.r.t. t_1, t_2): Divergence (data-histogram, best fit, over interval $[t_1 , t_2]$)(P)

By formulating and solving problem (P) one can find optimal threshold values t_1 and t_2 and accordingly the process of finding SE (and subsequently HE) could be "automated". This is a very interesting research problem where different ideas could be implemented. However, there are many difficulties in implementing this process. For example, the question of how to define "Divergence" turns out to be quite complicated. Taking this into account in [8], the best threshold values t_1 and t_2 for each output variable are defined "manually" by closer comparison of the distribution of outlying values and the best fit found in **Step 3**. In this paper, we propose the approach/algorithm described below.

2.4 Algorithm for Defining Interval Sub-Extreme (SE)

Also referred to hereafter as the SE divergence algorithm, let $h(i)$, $i \in \{1, 2, \ldots, N\}$ be the distribution function for a particular outlying measure (VLM in this case) and let $f(t)$ be the best fit found in Step 3.

SE Step-1: First we calculate a scaled difference between $h(i)$ and $f(t)$ defined as:

$$\text{Divergence } (h(i),f(i)) = \frac{h(i) - f(i)}{\max(f(i), M)} \text{ for all } i\, \varepsilon \{1, 2, \ldots, N\}$$

Here M is the minimal number-threshold value that is used to keep the same scaling for all the right tale where the best-fit $f(i)$ values approach to zero (accordingly the ration becomes infinitely large).

SE Step-2: Find the interval $[t_1^0, t_2^0]$ on the right tale where divergence $(h(i),f(i))$ values are overall larger than some given threshold. The corresponding optimization problem is formulated below:

Maximise: $\sum_{t=t_1}^{t_2}$ Divergence $(h(t),f(t))$

Subject to: $t_2 > t_1;t_1, t_2 \in [t^{min}, t^{max}]$

$$h(t_1) - f(t_1) > Thresh, h(t_2) - f(t_2) > Thresh$$

In this formulation, considering the sum as an objective function allows us to consider the case that does not require the inequality $h(t) > f(t)$ at all points of the interval $\left[t_1, t_2\right]$ which is crucial when dealing with histograms.

SE Step-3: Find the shortest interval $[t_1^*, t_2^*]$ in $[t_1^0, t_2^0]$ such that the sum of divergence $(h(i),f(i))$ over $[t_1^*, t_2^*]$ constitutes a significant portion of the sum of divergence $(h(i),f(i))$ over $\left[t_1^0, t_2^0\right]$. The corresponding optimisation problem is formulated below:

Minimise: $t_2 - t_1$

Subject to: $t_2 > t_1; t_1, t_2 \in [t_1^0, t_2^0]$

$$\sum_{t=t_1}^{t_2} \text{Divergence } (h(t),f(t)) > P * \sum_{t=t_1^0}^{t_2^0} \text{Divergence } (h(t),f(t))$$

Parameter Settings: Calculations show that 5–10% of the maximum of the best fit f in SE Step-1 provide almost the same results. Accordingly, in all calculations below we use 5% and set

$$M = Max(0.05 * Max(f), 1)$$

In SE Step-2, we use $\left[t^{min}, t^{max}\right] = [\frac{N}{3}, N]$ so that the interval that we are looking for is on the right-hand side of the distribution. Moreover, we set:

$$Thresh = mean\left(\text{Divergence } (h(t),f(t)), t \in \left[\frac{N}{2}, N\right]\right)$$

which provides an initial average value for the difference between the data-histogram and the best fit on the right tale. This allows us to deal with the case when there is a section on the right tale with "unusual" high values of $h(t)$ but the best fit $f(t)$ still lies above $h(t)$.

For the parameter P in SE Step-3, we consider different values including 0.9 and 0.95 and report corresponding intervals for each output variable.

3 Results

Consistent with previous analysis [8] sample data are derived from service level calculations of three health provider disciplines being Physiotherapy, General Practice and Psychiatry. Each discipline has resulted in a Modal and Specialised cluster for subsequent anomaly detection along with a component of the additional Aberrant cluster being applied to General Practice and denoted as High servicing. Each cluster has been derived using decomposition of related service level groupings as described below and relative to MBS or internal TAC service item definitions where appropriate.

3.1 Physiotherapy Service Levels

The first sample dataset contains aggregated Physiotherapy billing data of 31,447 health provider/client combinations and is representative of 396,472 underlying transactions over a 30 month period. Local neighbourhood in this instance has been set at the closest 100 points. For simplicity of labelling local distance observations are referred to as the variation from local mean (VLM).

Physiotherapy service levels are clearly a combination of distinct sub-populations noted by the multiple peaks in the related VLM histogram (Fig. 1). This is attributed to both specific service level preferences within provider sub-groups (e.g. Neurophysio extended head/spinal injury consultations or less expensive hydrotherapy/group sessions) and the nature of the initial encoding of these service levels (i.e. equivalent time-based encoding of consultations <30 mins, = 30 mins and >30 mins in duration). Service levels correspond to this time-based encoding and are referred to as service level 1, 2 and 3, respectively. For simplicity in data handling and analysis, the service level combinations are represented by a binary vector corresponding to the use of each of the 3 service levels for a particular provider–client combination. With 3 service levels, this gives a maximum of 7 valid combinations represented by the binary vectors 001, 010, 011, 100, 101, 110 and 111 (e.g. 001 represents service level 3 only, 010 represents service level 2 only, etc.).

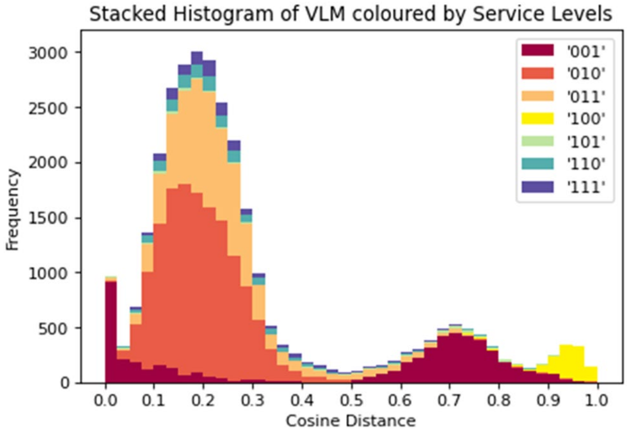

Fig. 1 Stacked histogram of Physiotherapy service levels (Cosine distance measure) using variation from local mean (VLM) [8]

To determine appropriate local anomaly thresholds, it is necessary to examine service level combinations in separate but related groupings by decomposition. Given service level 2 (= 30 mins) is the most prevalent service modality, all service levels containing level 2 are considered together (i.e. 010, 011, 110 and 111). Digression from service level 2 can be an indicator of over-servicing by a provider and warrants further investigation or review. The SE divergence algorithm results demonstrated in Fig. 2 return an interval of 0.580 and 0.719 with a P parameter of 0.90 and an interval of 0.560 and 0.719 with a P parameter of 0.95.

Service levels 001 and 101 with VLM > = 0.45 have more appropriate thresholds of divergence identified between cosine distance 0.86 and 0.92 when applying a minimally transformed normal distribution. The SE divergence algorithm results demonstrated in Fig. 3 return an interval of 0.886 and 0.923 with a P parameter of 0.90 and an interval of 0.875 and 0.923 with a P parameter of 0.95. The extreme nature and minimal business value of the remaining service level group, 100, is cause for its omission from further VLM anomaly calculations in this discipline.

It is clear that use of the cosine distance measure with this output domain leads to an abnormal distribution at the local level requiring appropriate consideration. More specifically there are two local sub-groups which benefit from a tailored threshold, the modal and the 001/101 service levels with VLM >= 0.45.

3.2 General Practitioner Service Levels

The second sample dataset used in this study contains aggregated General Practitioner billing data of 35,116 health provider/client combinations and is representative of 215,045 underlying transactions over a 24 month period. Local neighbourhood again has been set at the closest 100 points. The MBS defines the majority of General Practitioner consultations in four categories based on duration (Levels A, B, C and D). These categories, or levels, are prevalent across face-to-face consultations in a clinical setting, home visits and more recently via telehealth (telephone or video conferencing consultations). The levels are consistent regardless of the delivery type and correspond to the following: Level A–less than 6 minutes; Level B–6 to 20 minutes; Level C–21 to 40 minutes; Level D–greater than 40 minutes in duration.

General Practitioner service levels are also clearly a combination of distinct sub-populations noted by the multiple peaks in the related VLM histogram (Fig. 4). This is

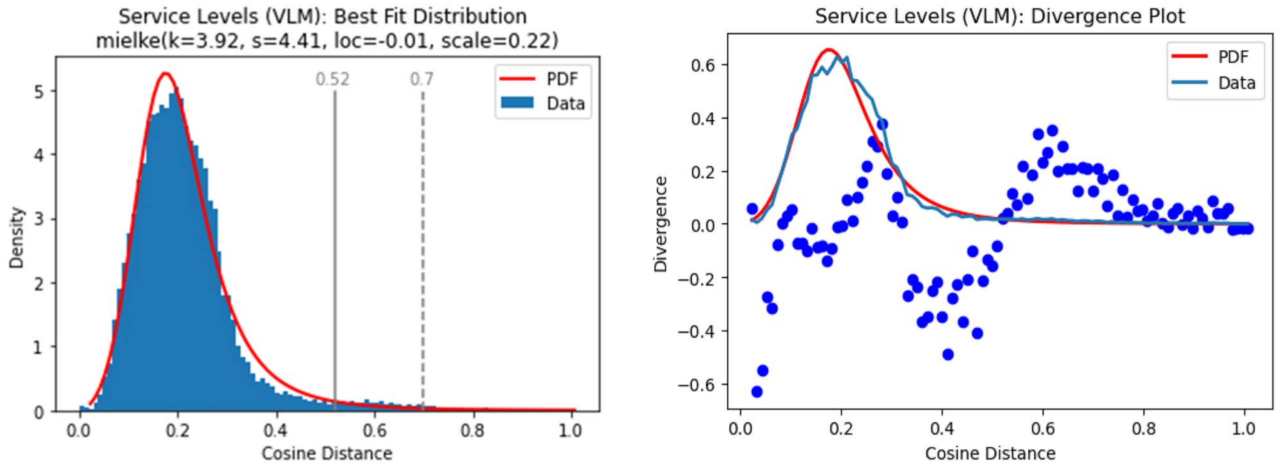

Fig. 2 Histogram of Physiotherapy service levels (Cosine distance measure) showing modality servicing (010, 011, 110 and 111) and best-fit distribution of VLM (PDF = probability density function of relevant distribution). Minor tail divergence evident with manual divergence interval determined at cosine distance values of 0.52 and 0.7 [8]. Divergence plot contrasting VLM data and best-fit PDF with divergence calculated as per SE divergence algorithm

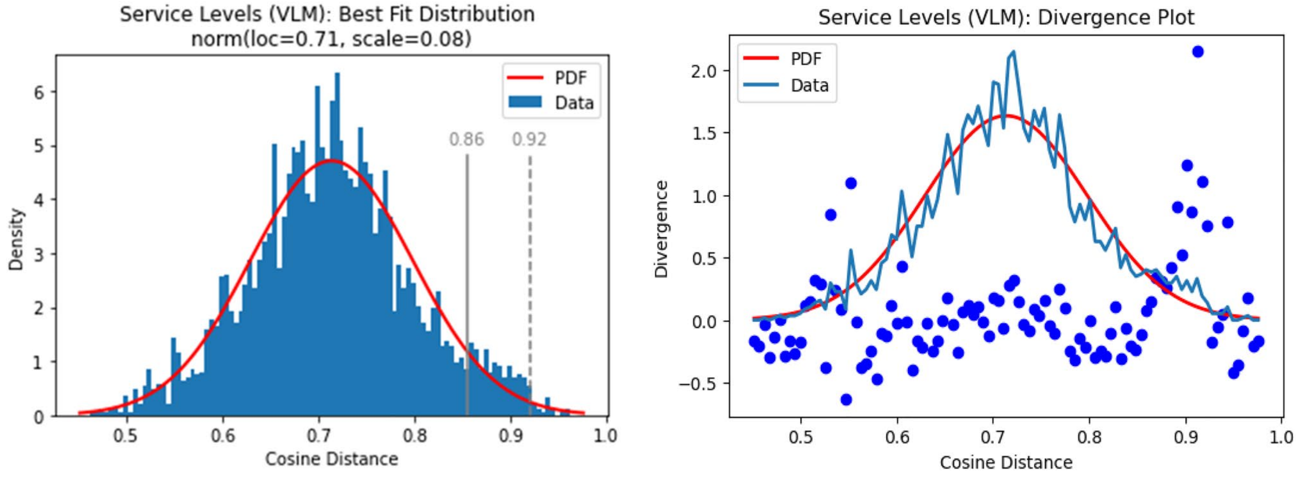

Fig. 3 Histogram of Physiotherapy service levels (Cosine distance measure) showing specialised servicing (001 and 101) and best-fit distribution of VLM (PDF = Probability Density Function of relevant distribution). Prominent tail divergence evident with manual divergence interval determined at cosine distance values of 0.86 and 0.92 [8]. Divergence plot contrasting VLM data and best-fit PDF with divergence calculated as per SE divergence algorithm

attributed to both specific service level preferences amongst providers (e.g. frequent use of Level B and Level C consults over Level A or Level D) and the nature of the initial encoding of these service levels (i.e. MBS-based encoding of consultations based on duration). Service levels correspond to this time-based encoding and are referred to as service level 1, 2, 3 and 4, respectively. With 4 service levels, the binary vector representation gives a maximum of 15 valid combinations represented by the binary vectors 0001, 0010, …, 1111 (e.g. 0001 represents service level D only, 0010 represents service level C only, etc.).

Defining appropriate local anomaly thresholds again requires examining service level combinations in separate but related groupings by decomposition. Service level B (0100) is the most prevalent service modality, hence all service levels containing level B are considered together along with the majority of level A (i.e. 0100, 0101, 0110, 0111, 1001, 1010, 1011, 1100, 1101, 1110 and 1111). The SE divergence algorithm results demonstrated in Fig. 5 return an interval of 0.478 and 0.603 with a P parameter of 0.90 and an interval of 0.478 and 0.626 with a P parameter of 0.95.

Likewise service levels 0010 and 0011 have more appropriate thresholds of divergence identified between cosine distance 0.79 and 0.84 when applying a minimally transformed normal distribution. The SE divergence algorithm results demonstrated in Fig. 6 return an interval of 0.797 and

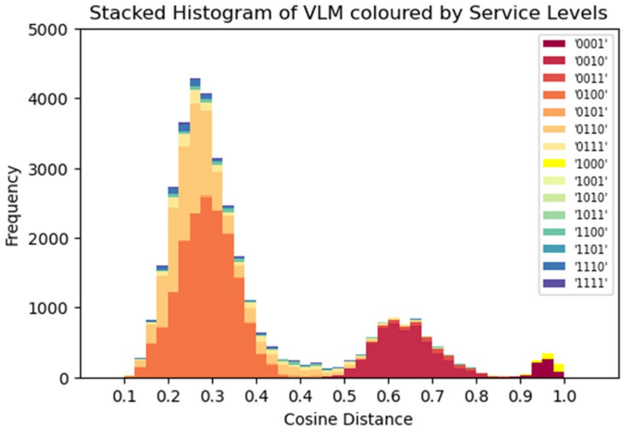

Fig. 4 Stacked histogram of General Practitioner service levels (Cosine distance measure) using variation from local mean (VLM) [8]

0.824 with a P parameter of 0.90 and an interval of 0.797 and 0.836 with a P parameter of 0.95.

Service Level 0001 on its own has more appropriate thresholds of divergence identified between cosine distance 0.976 and 0.985 when applying a power-normal distribution. The SE divergence algorithm results demonstrated in Fig. 7 return an interval of 0.975 and 0.986 with a P parameter of 0.90 or 0.95. The extreme nature and minimal business value of the remaining service level group, 1000, is cause for its omission from further VLM anomaly calculations as was the experience with Physiotherapy.

It is clear that use of the cosine distance measure with this input domain also leads to an abnormal distribution at the local level requiring appropriate consideration. Most interesting is that the number of service levels does not necessarily translate directly to the number of population sub-groups requiring specific analysis.

3.3 Psychiatric Service Levels

The third sample dataset used in this study contains aggregated Psychiatrist billing data of 3,363 health provider/client combinations and is representative of 34,087 underlying transactions over a 24-month period. Local neighbourhood again has been set at the closest 100 points. The MBS defines the majority of Psychiatric consultations in five categories based on duration (Levels A, B, C, D and E). These categories, or levels, are prevalent across face-to-face consultations in a clinical setting and more recently via telehealth. The levels are consistent regardless of the delivery method and correspond to the following: Level A–less than 15 minutes; Level B–15 to 30 minutes; Level C–greater than 30 to 45 minutes; Level D–greater than 45 to 75 minutes; Level E–greater than 75 minutes in duration.

Psychiatric service levels are also a combination of distinct sub-populations which becomes evident upon inspection of the stacked VLM histogram (Fig. 8). This is attributed to both specific service level preferences amongst providers (e.g. frequent use of Level B, C and D consults over Level A or Level E) and the time-based nature of the initial encoding. Service levels corresponding to this time-based encoding and are referred to as service level 1, 2, 3, 4 and 5, respectively. With 5 service levels the binary vector representation gives 24 valid combinations since 7 service level combinations are not present in the population.

Defining appropriate local anomaly thresholds again requires examining service level combinations in separate but related groupings by decomposition. Service level D

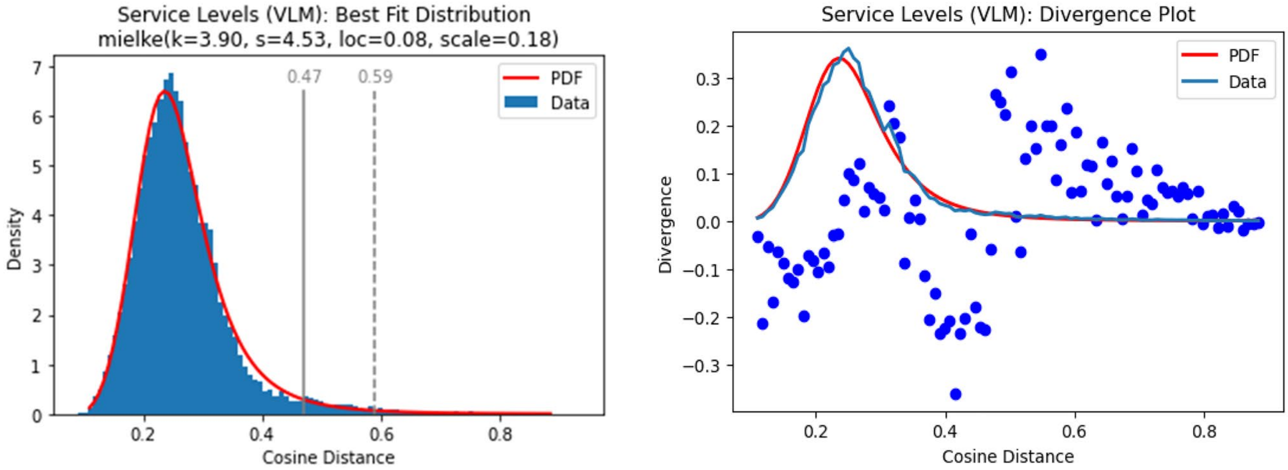

Fig. 5 Histogram of General Practitioner service levels (Cosine distance measure) showing modality servicing (0100, 0101, 0110, 0111, 1001, 1010, 1011, 1100, 1101, 1110 and 1111) and best-fit distribution of VLM (PDF = Probability Density Function of relevant distribution). Minor tail divergence evident with manual divergence interval determined at cosine distance values of 0.47 and 0.59 [8]. Divergence plot contrasting VLM data and best-fit PDF with divergence calculated as per SE divergence algorithm

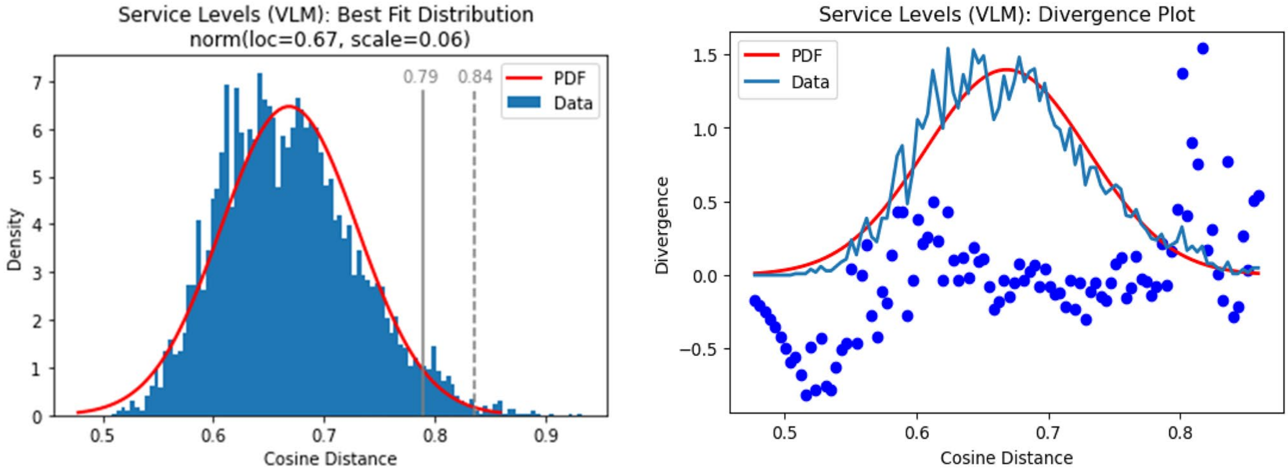

Fig. 6 Histogram of General Practitioner service levels (Cosine distance measure) showing specialised servicing (0010 and 0011) and best-fit distribution of VLM (PDF = probability density function of relevant distribution). Prominent tail divergence evident with manual divergence interval determined at cosine distance values of 0.79 and 0.84 [8]. Divergence plot contrasting VLM data and best-fit PDF with divergence calculated as per SE divergence algorithm

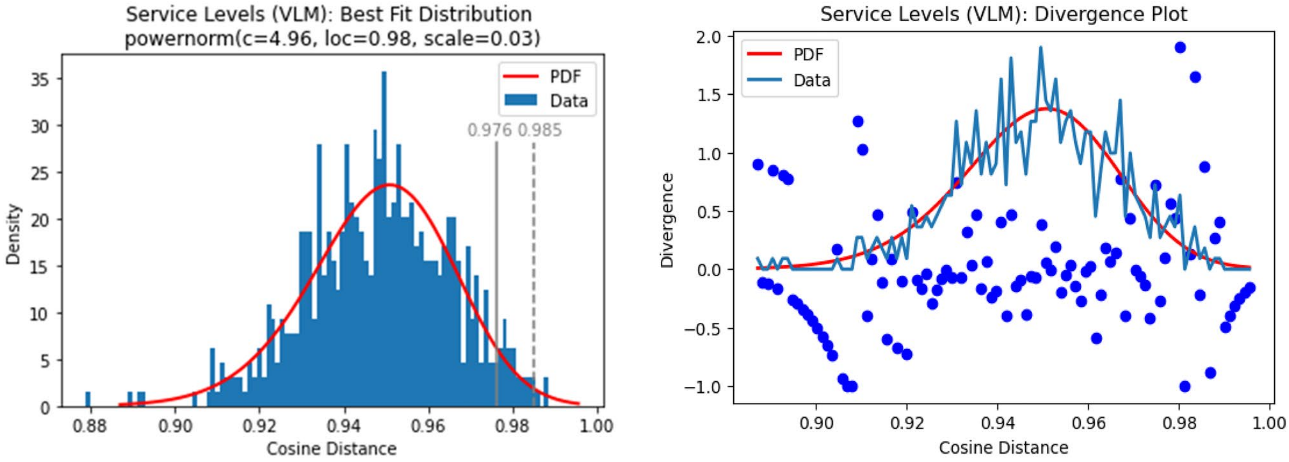

Fig. 7 Histogram of General Practitioner service levels (Cosine distance measure) showing high servicing (0001) and best-fit distribution of VLM (PDF = probability density function of relevant distribution). Moderate tail divergence evident with manual divergence interval determined at cosine distance values of 0.976 and 0.985 [8]. Divergence plot contrasting VLM data and best-fit PDF with divergence calculated as per SE divergence algorithm

(00010) is the most prevalent service modality followed by the combination of levels C and D (00110) and then level C only (00100). The modality service grouping remains after removing service levels 10000, 01000, 11000 and 00001. The SE divergence algorithm results demonstrated in Fig. 9 return an interval of 0.701 and 0.793 with a P parameter of 0.90 and an interval of 0.616 and 0.793 with a P parameter of 0.95.

Service levels 01000 and 11000 have more appropriate thresholds of divergence identified between cosine distance 0.895 and 0.935 when applying a power-normal distribution. The SE divergence algorithm results demonstrated in Fig. 10 return an interval of 0.921 and 0.924 with a P parameter of 0.90 or 0.95.

The extreme nature and minimal business value of the remaining service level group, 10000 and 00001, is cause for its omission from further VLM anomaly calculations.

3.4 SE Interval Summary by Discipline

Results from SE interval determinations are shown in Table 1 with associated record counts and threshold parameter values for P. In comparison with setting of the SE divergence interval manually [8], it is observed that the

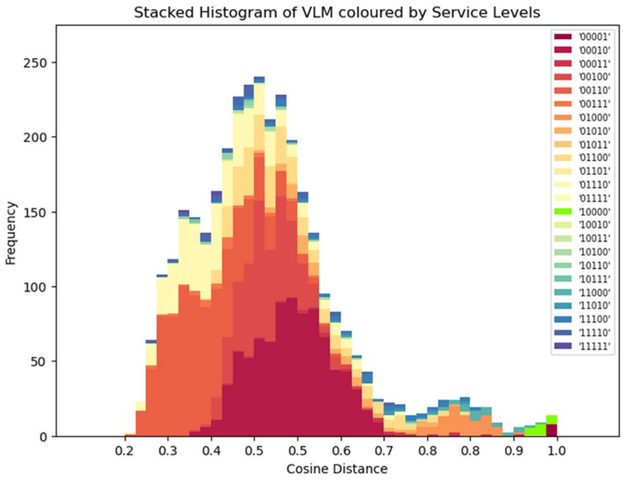

Fig. 8 Stacked histogram for Psychiatric service levels (Cosine distance measure) using variation from local mean (VLM) [8]

proposed algorithm, in general, returns a reduced interval width with increased sensitivity for the resultant limit values. This is particularly evident for modality servicing, where the majority of observations reside, and hence a reduction in anomalous observations can be expected with intervals derived using the SE divergence algorithm and a P value of 0.9. The reduction in SE observations, as reported in Table 1, equates to 256 provider/claims for Physiotherapy (36% reduction), 91 provider/claims for General Practice (10% reduction) and 19 provider/claims for Psychiatry (23% reduction). Tailoring of the P parameter gives control over the magnitude of the interval returned and can be matched to specific operational constraints as required.

3.5 Comparison with Existing Methods

In this section, we compare the aforementioned results using the same 26 unscaled input data elements for General Practitioner (GP) service levels (injury vector, age, time from accident and 4 service levels) with an existing benchmark outlier method. GP data are considered representative of all disciplines for the purposes of this exercise. The unsupervised anomaly detection method, Local Outlier Factor (LOF), was selected with a local neighbourhood parameter setting of 100 points. This method is also based on a local distance measure which is consistent with our approach. Given the provider–claim population of 35,116 observations a LOF score of − 1.5 was used to determine an equivalent number of overall anomalies. Initial results sorted by health provider ID (or observation number) are shown in Fig. 11 and confirm that variability in the multi-dimensional input space makes anomaly detection difficult based on raw input values alone.

An interesting commonality observed between LOF and SE/HE anomalies occurs amongst recently created provider IDs. Recent creation of a provider record on the source provider billing and payment system results in a higher ID number since new IDs are allocated sequentially. This is consistent with the final output ranking results for General Practitioners where aggregated anomalies showed newly created doctors providing post-operative care through the use of higher level, home visit, service items were divergent from the overall GP population in their billing behaviour.

A more accurate demonstration of SE/HE anomalies versus LOF in this discipline is evident when observations are resorted by the local service level cosine distance calculation as in Sect. 3.2 (see Fig. 12).

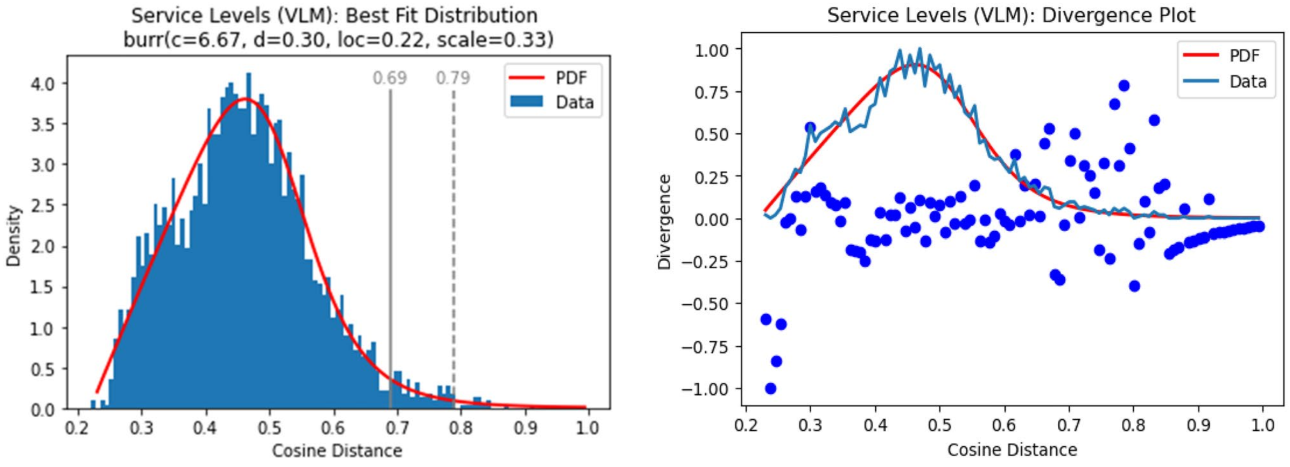

Fig. 9 Histogram of Psychiatric service levels (Cosine distance measure) showing modality servicing (predominantly Levels C and D) and best-fit distribution of VLM (PDF = probability density function of relevant distribution). Moderate tail divergence evident with manual divergence interval determined at cosine distance values of 0.69 and 0.79 [8]. Divergence plot contrasting VLM data and best-fit PDF with divergence calculated as per SE divergence algorithm

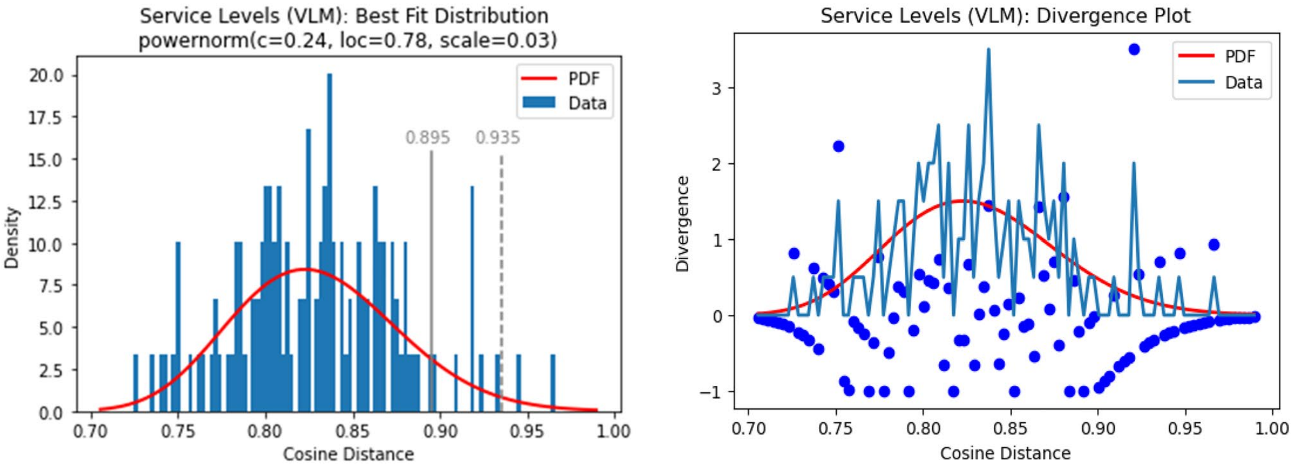

Fig. 10 Histogram of psychiatric service levels (Cosine distance measure) showing specialised servicing (Level B with Level A/B, 01000 and 11000) and best-fit distribution of VLM (PDF = probability density function of relevant distribution). Minor tail divergence evident with manual divergence interval determined at cosine distance values of 0.895 and 0.935 [8]. Divergence plot contrasting VLM data and best-fit PDF with divergence calculated as per SE divergence algorithm

Table 1 Sub-extreme interval determinations for Physiotherapy, General Practice and Psychiatry disciplines: Modal, Specialised and High clusters (where applicable)

			Interval [t_1^*, t_2^*]					
Discipline	Cluster	Obs	From [8]	Obs	$P = 0.95$ (SE Step-3)	Obs	$P = 0.90$ (SE Step-3)	Obs
Physiotherapy	Modal	24,378	[0.52, 0.70]	512	[0.560, 0.719]	410	[0.580, 0.719]	351
	Specialised	3,752	[0.86, 0.92]	198	[0.875, 0.923]	138	[0.886, 0.923]	103
General Practice	Modal	28,642	[0.47, 0.59]	685	[0.478, 0.626]	704	[0.478, 0.603]	649
	Specialised	5,668	[0.79, 0.84]	183	[0.797, 0.836]	141	[0.797, 0.824]	124
	High	586	[0.976, 0.985]	27	[0.975, 0.986]	31	[0.975, 0.986]	31
Psychiatry	Modal	3,212	[0.69, 0.79]	76	[0.616, 0.793]	235	[0.701, 0.793]	64
	Specialised	124	[0.895, 0.935]	8	[0.921, 0.924]	1	[0.921, 0.924]	1

Observation counts show parameter value $P = 0.9$ reduces sub-extreme anomalous observations by 36% in Physiotherapy, 10% in General Practice and 23% in Psychiatry data

The distribution of GP service levels indicates that LOF outliers are scattered throughout the population. This leads to difficult interpretation from a business perspective when defining the causal factors behind certain groups of anomalies. Appropriate scaling of inputs and decomposition of service level groupings into Modal, Specialised and Aberrant clusters before defining intervals of SE and HE divergence is an appropriate approach in TAC's context to preserve business knowledge of the source data and aid resultant interpretation of anomalous results.

4 Summary

Detailed considerations undertaken in this study enhance the detection capability of the original framework by standardising the interval determination for sub-extreme anomalies. Whilst manual threshold setting was considered appropriate previously, applying algorithmic rigour has produced more appropriate limits which achieve the final goal of targeting minimal provider–claim combinations for review. From a business perspective less observations flagged as anomalous ultimately translates into less cases for review and intervention, saving valuable time and resources when actioning model output. Comparison with a contemporary outlier detection method (LOF) was undertaken and produced two distinct findings. Firstly that agreeance exists between the two methods when recently created providers exhibit behaviours which are divergent from their peer group. Secondly that the introduction of domain knowledge regarding appropriate service level groupings aids the final interpretability of anomalous results. Limitations are inherent in a framework which relies on deep domain knowledge for its definitions and subsequent output translation, particularly when it

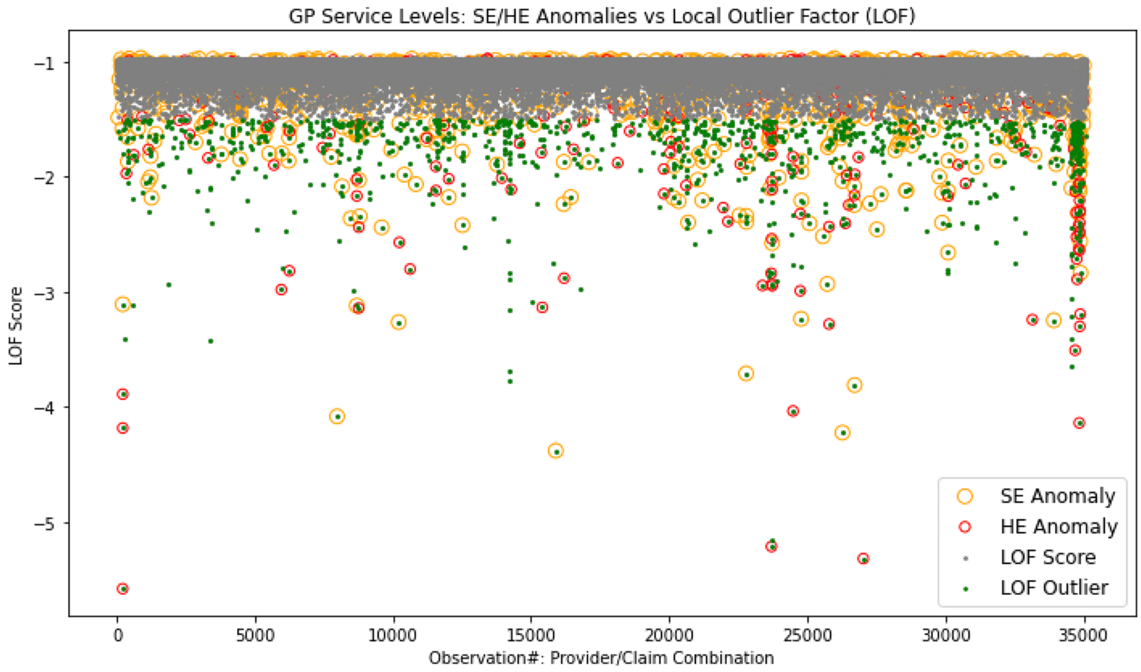

Fig. 11 Overlaid scatterplot of sub-extreme/high extreme anomalies versus Local Outlier Factor (LOF) scores with a LOF threshold of − 1.5 for General Practitioner (GP) service level input data. Observations sorted by health provider ID indicate a general spread of anomalies/outliers across the entire domain for both methods. An interesting observation is the commonality amongst recently created providers with higher ID values being evident at the right extremity

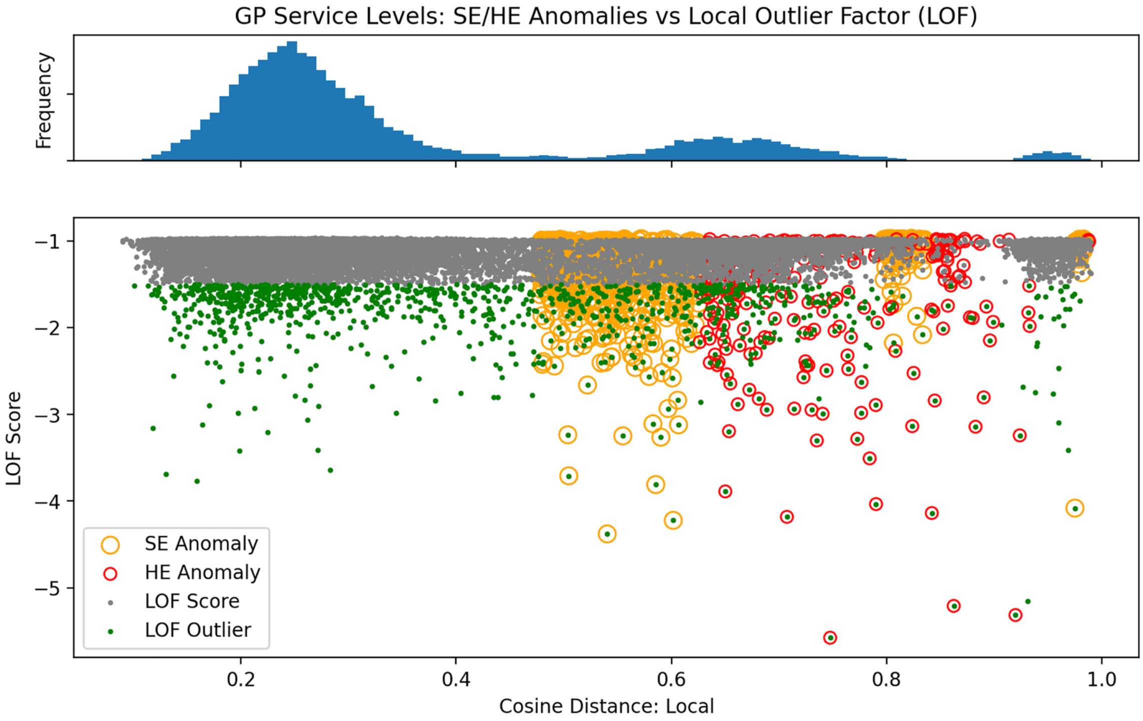

Fig. 12 Overlaid scatterplot of sub-extreme/high extreme anomalies versus Local Outlier Factor (LOF) scores with a LOF threshold of −1.5 for General Practitioner (GP) service level input data. Observations sorted by local service level cosine distance values to coincide with frequency distribution used prior to decomposition (see Fig. 4). LOF outliers are scattered throughout the distribution which in turn makes the business interpretation of these results difficult

comes to maintenance, reuse and indeed portability, and this method is no different in that respect. Anomaly detection algorithms which rely on extreme values, of which there are many, are not well suited to this specific business problem where sub-extreme anomalies are sought for their potential in flagging health provider billing behaviours which “push the envelope” amongst their peers.

Declarations

Ethical Approval The primary/corresponding author is a substantive employee of the custodian organisation (Transport Accident Commission of Victoria) in addition to being a research degree student at Deakin University. The authors have no additional interests to declare that are relevant to the content of this article. Data have been approved for use in a research capacity by the TAC and is fully de-identified to protect individual and health provider privacy. This study has received Deakin University ethics approval (reference number: 2021-363).

References

1. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3):1–58
2. Kirlidog M, Asuk C (2012) A fraud detection approach with data mining in health insurance. Procedia-Social Behavioral Sci 62:989–94
3. Koh H, Tan G (2005) Data mining applications in healthcare. J Healthcare Inf Manage: JHIM 19(2):64–72
4. Kumaraswamy N, Markey MK, Barner JC, Rascati K (2022) Feature engineering to detect fraud using healthcare claims data. Expert Syst Appl 210:118433
5. Lu J, Fung BC, Cheung WK (2020) Embedding for anomaly detection on health insurance claims. In: 2020 IEEE 7th international conference on data science and advanced analytics (DSAA), pp 459-68
6. Mammadov, M, Muspratt, R, Ugon, J (2017) Detection of outlier behaviour amongst health/medical providers servicing TAC clients. In: AusDM 2017: Proceedings of the 15th Australasian conference on data mining, vol 845, pp. 161-72.
7. *Medicare Benefits Schedule* 2021, Australian Government Department of Health.
8. Muspratt R, Mammadov M (2022) Decomposition of service level encoding for anomaly detection. In: Data mining: 20th Australasian conference, AusDM 2022, Western Sydney, Australia, December 12–15, 2022, Proceedings, pp. 192-204
9. Omar A, Raed A, Terence S, Xiaogang M (2020) A review of local outlier factor algorithms for outlier detection in big data streams. Big Data Cognit Comput 5(1):1
10. Peng Y, Yang Y, Xu Y, Xue Y, Song R, Kang J, Zhao H (2021) Electricity theft detection in AMI based on clustering and local outlier factor. IEEE Access 9:107250–9
11. Shin H, Park H, Lee J, Jhee WC (2012) A scoring model to detect abusive billing patterns in health insurance claims. Expert Syst Appl 39:7441–50
12. Smiti A (2020) A critical overview of outlier detection methods. Comput Sci Rev 38:100306
13. Suboh S, Aziz, IA (2020) Anomaly detection with machine learning in the presence of extreme value-A review paper. In: 2020 IEEE conference on big data and analytics (ICBDA), pp. 66-72
14. Thomas A, Clémençon S, Gramfort A, Sabourin A (2017) Anomaly detection in extreme regions via empirical mv-sets on the sphere. In: artificial intelligence and statistics, pp 1011-9
15. Xu H, Zhang L, Li P, Zhu F (2022) Outlier detection algorithm based on k-nearest neighbors-local outlier factor. J Algor Comput Technol 16:1–12
16. Xu X, Liu H, Yao M (2019) Recent progress of anomaly detection. Complexity. https://doi.org/10.1155/2019/2686378