Application of Kohonen Maps in Predicting and Characterizing VAT Fraud in Southern Mozambique

Abstract

With the massive expansion of mobile information and communication technologies in sub-Saharan Africa, tax administrations today face even greater challenges to curb the Value Added Tax (VAT) fraud and tax evasion. Since its inception, in 1999, VAT has contributed to the largest share of Mozambique's tax revenues, but still has a relatively modest tax efficiency if compared to continental standards. This trend can be reversed by strengthening audit and inspection processes of the Mozambique Revenue Authority (MRA) with Data Mining, taking advantage of historical data stored by different information systems of the MRA. A Case Study of the southern region of Mozambique is presented, where some historical data available from tax audits are compared with the VAT returns using Kohonen Maps. Comparing the experimental results with other anomaly detection algorithms, Kohonen maps prove to be of great value in predicting and characterizing VAT fraud in Mozambique.

Keywords: Mozambique – VAT – Fraud – Audit – Data Mining – Kohonen Maps

1. Introduction

Mozambique introduced Value Added Tax in 1999 by Law No. 3/98, of 8 January, along with the respective VAT Code, approved by Decree No. 51/98, of 29 September. Despite of being very similar to the Portuguese, Mozambique VAT has specific adaptations driven by a protectionist regime for informal economy and tax exemptions in the areas of agriculture and fisheries (Palma, 2015: 379; 387-391).

Historically, VAT collections do have always-greater outputs in the southern region of Mozambique[1], with the province and city of Maputo together, being accountable for roughly 90% of VAT revenues.

Consequently, VAT constitutes as a barometer (Hajdúchová *et al.*, 2015: 676-681) to measure a country's Fiscal Gap (Hutton, 2017: 3-25), which is the estimate that results from the difference between the volume of expenditure and debt of a country and the volume of its revenue collection over a period of time, assuming a negative Fiscal Gap as non-existent.

Considered as a very common insufficiency in Sub-Saharan Africa (Krever, 2008:71-80), the Fiscal Gap of Mozambique is rooted in objective (La Feria and Schoeman, 2019: 961-962) and subjective (Manjate, 2018: 54-57) causes, and directly proportional to the VAT fraud rate (La Feria and Schoeman, 2019: 953 – 957; Portela, 2014: 7 – 28).

According to a study by UNICEF (2019) on Mozambique, a 30% increase[2] in VAT collection of internal and external trade would enable the elimination of Fiscal Gap by 2022, reaching a positive inflow of -

[1] *cf.* World Bank (2019), here: http://pubdocs.worldbank.org/en/379961580834119883/Mozambique-December-2019-Data-Capsule.xls (accessed on 25/05/2021).
[2] Estimate prior to the COVID-19 pandemic and the 2020-21 terrorist acts in Cabo Delgado.

0.8% to GDP[3], *i.e.,* 1.8% more than the baseline scenario used in this forecast. Furthermore, with the resulting surplus, the external debt could even reach 91.8% of GDP by 2024.

Mozambique could dramatically reduce its Fiscal Gap with a more efficient VAT management[4]. In consequence, Mozambique Revenue Authority (MRA) has invested since 2006 in the technological modernization of information systems (PARPA II, 2009: 29-31; 76-77) to interconnect with other relevant systems of the Government and the Private Sector.

As a result, greater transparency and better management of all tax processes are now being achieved, which has ultimately resulted in a generation of large volumes of structured and unstructured tax data from invoicing systems, particularly those related to VAT.

The baselines are now launched towards the Big Data analysis of Mozambican tax ecosystem, in line with the same stance followed by other contemporary tax administrations (IMF, 2017), further strengthening the need for Knowledge Discovery, to build a more assertive Taxpayer profile, despite being a relatively new field of study in the Mozambican public and government sector (Sotomane, 2014: 102-103).

2. Case Study

The Case Study consists of the analysis of VAT and Corporate Income Tax[5] returns of 27,049 taxpayers for the years 2013-2018. These data are from the information systems of the MRA and subjected to prior screening, particularly for the treatment of the numerous missing fields, which usually characterize VAT and income tax returns.

The sample population of Case Study are Normal VAT taxpayers referred to in Art. 19 of the VAT Code, whose refund is paid in full by the Government and also Normal VAT taxpayers whose inspections culminate in court cases, excluding those prosecuted by drugs, human trafficking and related offences, which are managed separately as criminal according to Mozambican Law.

The Case Study can be considered brand new in the context of Mozambique – and in some countries whose fiscal reality is similar (Mwanza and Phiri, 2016: 793-798; de Roux *et al.*, 2018: 215-222; Jihal *et al.*, 2018: 90-92; Jupri and Sarno, 2020: 75-87) – and is meant to answer three (3) research hypotheses: (i) the cases under investigation by the Inspection sector correlate with VAT frauds; (ii) refund payments correlate with potential VAT frauds; and (iii) the evaded VAT revenue far exceeds that recovered by the MRA.

[3] Gross Domestic Product, *i.e.*, the monetary value of all finished goods and services made within a country during a specific period.
[4] VAT Fiscal Gap is influenced by: (i) compliant statements; (ii) non-conforming statements; and (iii) undelivered statements.
[5] *i.e.,* Imposto sobre o Rendimento de Pessoas Colectivas (IRPC).

3. Kohonen Maps

Kohonen Maps (1982) or Self-Organized Maps are unsupervised artificial neural networks, which combine multidimensional clustering with visual techniques:

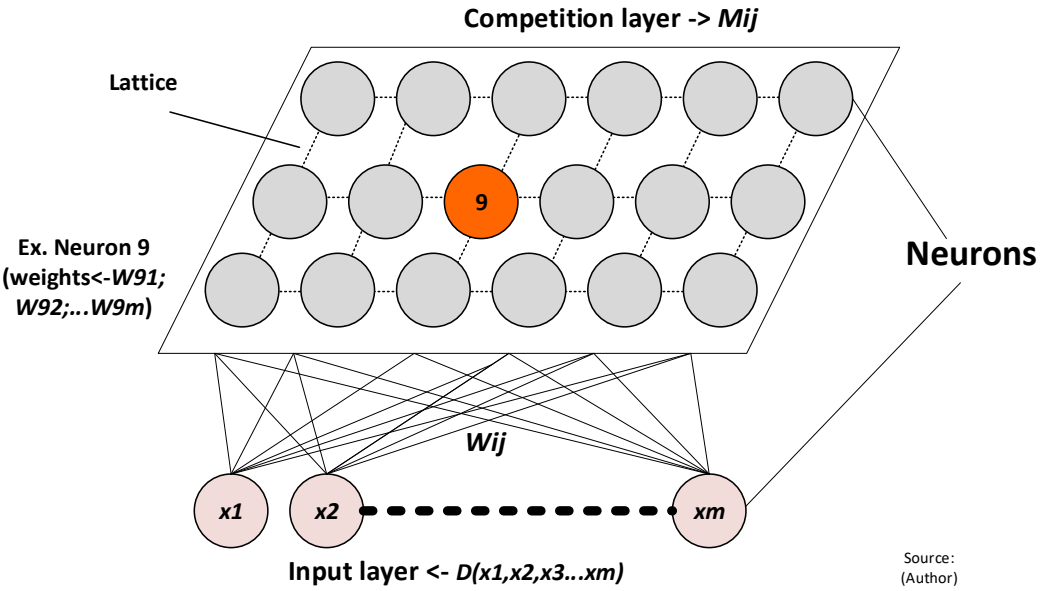

Figure 1 - Kohonen Map Architecture

The Kohonen Map architecture (Figure 1) is characterized by: (i) an *input layer*, composed of a data vector (or register) $D = (x_1 x_2 x_3 \dots x_m)$ and (ii) a *competition layer*, consisting of a matrix $M_{(i,j)}$, typically two-dimensional lattice[6] of neurons, which is geometrically ordered at each iteration, according to: (i) similarity measures, *e.g.* the Euclidean distance; and (ii) competitive[7] learning (Duda *et al.* 2000: Cap. 10.11). Since the optimization of the dimensions of $M_{(i,j)}$ is done by heuristics (Wendel and Buttenfield, 2010), it results in the empirical relationship

$$m = 5 \times \sqrt{n} \quad (1)$$

where m is the product of the dimensions of $M_{(i,j)}$ and n the number of observations – or records of the data sample. The functioning of Kohonen Maps explains in five steps (Kohonen, 1998: 159-167; Tu *et al.*, 2016: 1-6):

First, after normalizing[8] the vector $D = (x_1 x_2 x_3 \dots x_m)$ with any standard method, randomly initialize the weights W_{ij}.

Second, read $D = (x_1 x_2 x_3 \dots x_m)$ by the input layer and calculates the Euclidean distance[9] between $D = (x_1 x_2 x_3 \dots x_m)$ and the weights W_{ij}, *i.e.*

[6] Can be rectangular or hexagonal.
[7] In contrast to the multilayer perceptron and other neural networks, whose learning is done by minimizing errors.
[8] Not always necessary, as initialization can be done randomly with the original data.
[9] Other types of distance, such as Manhattan, Minkowski, Cosine, among many, could also be used.

$$d_j = \|D - W_{ij}\| = \sqrt{\sum_{i=1}^{m} \left(d_i(t) - w_{ij}(t)\right)^2} \quad (2)$$

where w_{ij} is the weight of the i input layer neurons and the j neurons of the competition layer. The winning neuron[10] BMU - Best Matching Unit - is obtained when $d_{BMU} = \min(d_j)$ is verified.

Third, determine the neighbourhood radius[11] of the neuron BMU with the interpolation function[12]

$$N_{BMU}(t) = N_0 e^{-\frac{t}{\lambda}} \ : N_{BMU} \to 0, t \to T \quad (3)$$

where $N_{BMU}(t)$ is the neighbourhood radius at training time$t$; N_0 the initial radius of the neighbourhood; and $\lambda = \frac{K}{\log(N_0)}$ a constant; and T the number of epochs. This interpolation function will shrink the neighbourhood radius as the algorithm progresses.

Fourth, update the weights of j in the competition layer

$$w_{ij}(t+1) = w_{ij}(t) + \eta(t)\left(d_i(t) - w_{ij}(t)\right) N_{BMU}(t) \ : 0 < \eta(t) < 1, \eta \to 0 \ se \ t \to T \quad (4)$$

where $\eta(t)$ is the learning rate; $N_{BMU}(t)$ neighbourhood radius; w_{ij} the weights of the neurons; and T the number of epochs.

Fifth, the output o_{BMU} of the first iteration is calculated and the halt condition is verified, *i.e*.

$$o_{BMU} = f\left(min_j \|D - W_j\|\right) \quad (5)$$

where *f* is an arbitrary activation function bounded by restrictions: (i) if $t \to T$ ends the algorithm, (ii) if not, repeat the cycle from the second to the fifth step.

The good performance of Kohonen Maps is highly influenced by the calibration of the number of epochs, neighbourhood function radius and learning rate essentially. In addition, two important metrics also contribute to this (Kohonen *et al.*, 2001; Tu *et al.*, 2016: 1-6): (i) the quantization error (QE) and (ii) the topographical error (TE). The metric (QE) determines the accuracy of data logging in the input layer as follows

$$QE = \frac{1}{T}\sum_{t=1}^{T} \|x(t) - w_k(t)\| \quad (6)$$

where $x(t)$ is a sample of $D = (x_{1,}x_2x_{3,} \dots x_{m,})$ at training time t; $w_k(t)$ is the weight associated to $x(t)$; and T the number of epochs. Therefore, the lower the QE the better the Kohonen Map learning rate, *i.e.,* clustering. Nevertheless, overfitting should always be avoided (Wendel and Buttenfield, 2010). In turn, the TE metric evaluates the consistency of the competition layer topography, by comparing the total number of data records that have two non-adjacent k_1 and k_2 neurons, which corresponds to

[10] Abbreviated BMU.
[11] Radius of the centremost BMU that *scans* adjacent neurons enabling their simultaneous update.
[12] Alternatively, neighbourhood functions such as Gaussian or Quadratic (bubble) could also be used.

$$TE = \frac{1}{T}\sum_{t=1}^{T} d(x(t)) \quad (7)$$

where $x(t)$ is a sample of $D = (x_1, x_2 x_3, \ldots x_m)$ at training time t; and is verified if $d(x(t)) = 1$, if k_1 and k_2 are not adjacent, or $d(x(t)) = 0$, otherwise. Consequently, the lower the TE the better the topographical quality of the Kohonen Map, *i.e.,* better visualization/projection of the results. For this reason, it is good practice to train as many Kohonen Maps as possible, using random parameterization, and then choose the one that minimizes QE and TE the most.

Kohonen maps have application in various branches of finance (Serrano-Cinca, 1998; Kiviluoto and Bergius, 1998) but also, in the detection of anomalies (Tian *et al.*, 2014), as a result of the combination with *k-Means* clustering (Steinhaus, 1957; Macqueen 1967) and KNN classification (Fix and Hodges, 1951) algorithms, originating a robust method with performance equivalent to that of other anomaly detection algorithms, such as Local Outlier Factor (Breunig *et al.*, 2000) and One-Class SVM (Schölkopf *et al.*, 1999: 582-588).

4. Methodology

To prove the research hypotheses, a six-stage methodology is observed: (i) extracting bulk data from various tax information systems; (ii) data validation using key fields; (iii) tagging of data associated with cases of fraud, investigations and suspicions; (iv) construction of the data cube used in Kohonen Maps; (v) proof of the investigation hypotheses; and (vi) analysis of the fraud behaviour patterns:

First, bulk data from: (i) 925,070 Normal VAT returns (2010-2019); (ii) 452,044 customs declarations (2018-2019); and (iii) 67,554 IRPC statements (2015-2019) are crosschecked by valid/existent Taxpayer Identification Numbers (TIN). Since Normal VAT returns are monthly based and IRPC statements annual, VAT returns are summed by fiscal year. Then after crosscheck bulk data with samples of audits, inspections and refunds provided by the MRA, namely: (i) 131 audits (2013-2018); (ii) 1,050 inspections (2018-19); and 660 VAT refund requests (2013-2019).

Second, from extracted data generated a valid sample of 27,049 taxpayers under the normal VAT and IRPC regime (2013-2018) for the Case Study, which represents about 50% of the tax population for that period. Then cross-check TIN, fiscal year and status of compliance to validate samples provided by the MRA, which results are: (i) 120 positive cases out of 131 audits[13]; (ii) 127 inspections of taxpayers out 1,050 cases submitted to the Fiscal Court[14]; and (iii) 85 refund payments out of 660 VAT refund requests[15].

Third, samples of audits, inspections and refunds provided by the MRA are successively marked as fraud (F), for audits with positive cases of fraud; investigations (I), for inspections submitted to the Fiscal Court; and suspicious (S), for VAT refunds actually paid. Since some audits are interspersed in different fiscal years, namely for 2013-2017; 2014-2018; and 2017-2018, it generates three partial results to be

[13] Approximately 0.44% of the taxpayers covered by the Case Study, among which 91.6% are positive cases of fraud, typically biased dataset as referred by Guarascio (2010).
[14] Samples from other VAT/ISPC regimes or with invalid/non-existent TIN were excluded.
[15] Samples with invalid/non-existent TIN were excluded.

compared with the consolidated map of the Case Study (2013-2018). Finally, samples of fraud (F), investigations (I) and suspicions (S) are interspersed, by economic sectors, among those 27,049 taxpayers of the Case Study (2013-2018), as summarized in Table 9:

Table 1 - Historical data on frauds, investigations and suspicions (2013-2018)

Economic sector	Incidence of fraud	Total of Taxpayers	F	I	S
A	Low	594	2	1	23
B	**High**	**10,276**	**49**	**93**	**21**
C	Medium-low	3,560	18	4	6
D	Low	3,756	7	5	8
E	Low	117	2	0	9
F	**Medium-high**	**7,981**	**39**	**22**	**13**
G	Low	765	3	2	5
	Total:	27,049	120	127	85

Fourth, a data cube (Wu *et al.*, 2012: 8769–8777) is created from Table1 with these characteristics: (i) zero value in all unfilled fields; (ii) mean value added to all fields with values equal to zero to suppress the error of dividing by zero; (iii) fiscal ratios (Basta *et al.*, 2009: 7-12; Vanhoeyveld *et al.*, 2019); and (iv) Z-Score normalization of fiscal ratios.

In the creation of tax ratios, the current weighting criteria of regular or unexpected audits by the MRA are taken into account: (i) fluctuation in sales volume; (ii) systematic credit; (iii) systematic losses; (iv) requests for refunds; (v) credit offset requests; (vi) anonymous or informant reporting; and (vii) financial behaviour of branches and sectors of activity. As there is no consolidated data on complaints and the financial behaviour of branches and sectors of activity, these are excluded criteria from the Case Study.

Fifth, apply Kohonen Maps to the Case Study samples (2013-2018) and record the frequency of frauds, investigations and suspicions. Validate Hypotheses 1 and 2 in this step. Finally, adopt the Mittal *et al.* (2018) principle for the treatment of truly positive cases[16] of fraud such as: (i) criminal cases submitted to the Public Prosecutor's in 2018 and 2019; and (ii) refund requests deferred and paid from 2014 to 2019. Validate Hypothesis 3 here as follows:

$$R_s = P_{\hat{Y}}.C_s \quad (8)$$

where $R_s; P_{\hat{Y}}; C_s$ are, respectively, the evaded revenue; the precision of the anomaly detection algorithm; and Mozambique's tax efficiency ratio[17], which stands at 35% (La Feria and Schoeman, 2019: 961-962), *i.e.,* evaded revenue represents close to 2.85 more than the volume of VAT currently submitted by taxpayers.

Sixth, apply Kohonen Components Plan to characterize behavioural pattern of the economic sectors that have historically presented the highest incidence of VAT fraud[18] using the consolidated map of the

[16] Which states that once registered as a positive case of fraud, the Taxpayer tends to relapse into fraudulent practices in future fiscal years.
[17] *cf.* S. Cnossen, "Mobilizing VAT Revenues in African Countries" (2015) *International Tax and Public Finance* 22(6), pp. 1077-1108.
[18] *cf.* Table 1 - Historical data on fraud, investigations and suspicions (2013-2018).

Case Study (2013-2018) as a reference. The same procedure applies to the most neglected economic sectors in homologous period.

5. Results

Querying the Case Study data reveals a strong correlation between some fiscal ratios, which is in line with the conclusions of recommended literature (Serrano-Cinca, 1998; Matos *et al.*, 2015). Consequently, 19 out of 23 initial fiscal ratios are chosen with the variable selectors with Decision Trees (CART), and only six (6) with Fisher's Discriminant Analysis. Here you are (Table 2) the comparative results of two Kohonen Maps anomaly detection methods with the Local Outlier factor (Breunig *et al.*, 2000) and the One-Class SVM (Schölkopf *et al.*, 1999:582-588):

Table 2 - Comparison of the performance of anomaly detection algorithms

Algorithm	Accuracy	Error rate	Recall	Specificity	Precision	F1-Score
Local Outlier Factor	99.11	0.89	0.00	99.55	0.00	0.50
One-Class SVM	98.45	1.55	3.13	99.59	8.33	0.52
Kohonen QE[19]	96.88	3.12	5.42	99.71	36.67	0.54
Kohonen KNN[20]	99.77	0.23	**85.37**	99.81	**58.33**	**0.85**

Overall, with 85.37% recall and 58.33% precision, Kohonen KNN with dimensionality reduced by Fisher's outperforms other anomaly detection algorithms. In any case, the performance of the majority of tested algorithms significantly exceeds 0.44% precision achieved by the random method of the MRA, which actually corresponds to 120 cases of fraud out of 27,049 taxpayers under the normal VAT regime.

For the characterization of the frauds (F), we follow a four-step methodology, considering only the reduced dimensionality with variable selectors with Decision Trees to preserve the majority of fiscal ratios of the Case Study.

It is spotted (Figure 2) that the U Matrix (Ultsch, 2003) has two distinct hemispheres: (i) the left, where the dark blue outline is visible, which denotes the minimization of the distances of the sample data; and (ii) the right hemisphere where, due to the maximization of the distances of the sample data, the presence of possible outliers is spotted. The Hits Map also supports this impression.

By clustering the neurons of the Kohonen Map with the k-Means, four clusters are obtained (Figure 2), which are topographically distributed over the same hemispheres, in the following order: (i) the left hemisphere, where a well-formed cluster 0 groups neurons with sample data with shorter intra-cluster

[19] Kohonen QE uses the traditional quantization error measurement approach. In this case, the 97th percentile is considered as a threshold.

[20] Kohonen KNN (Tian et al., 2014) previously removes neurons contaminated with noisy data and optimizes quantization errors with the density function of the log-normal distribution, with the 99.7 percentile as a threshold.

distances; on contrary (ii), in the right hemisphere, clusters 1, 2 and 3 isolate neurons with greater intra-cluster distances, likely outliers or ill-formed clusters.

Matching between the positive cases of the Kohonen Map and k-Means clustering (Figure 2) is achieved with the KNN classifier (Fix and Hodges, 1951):

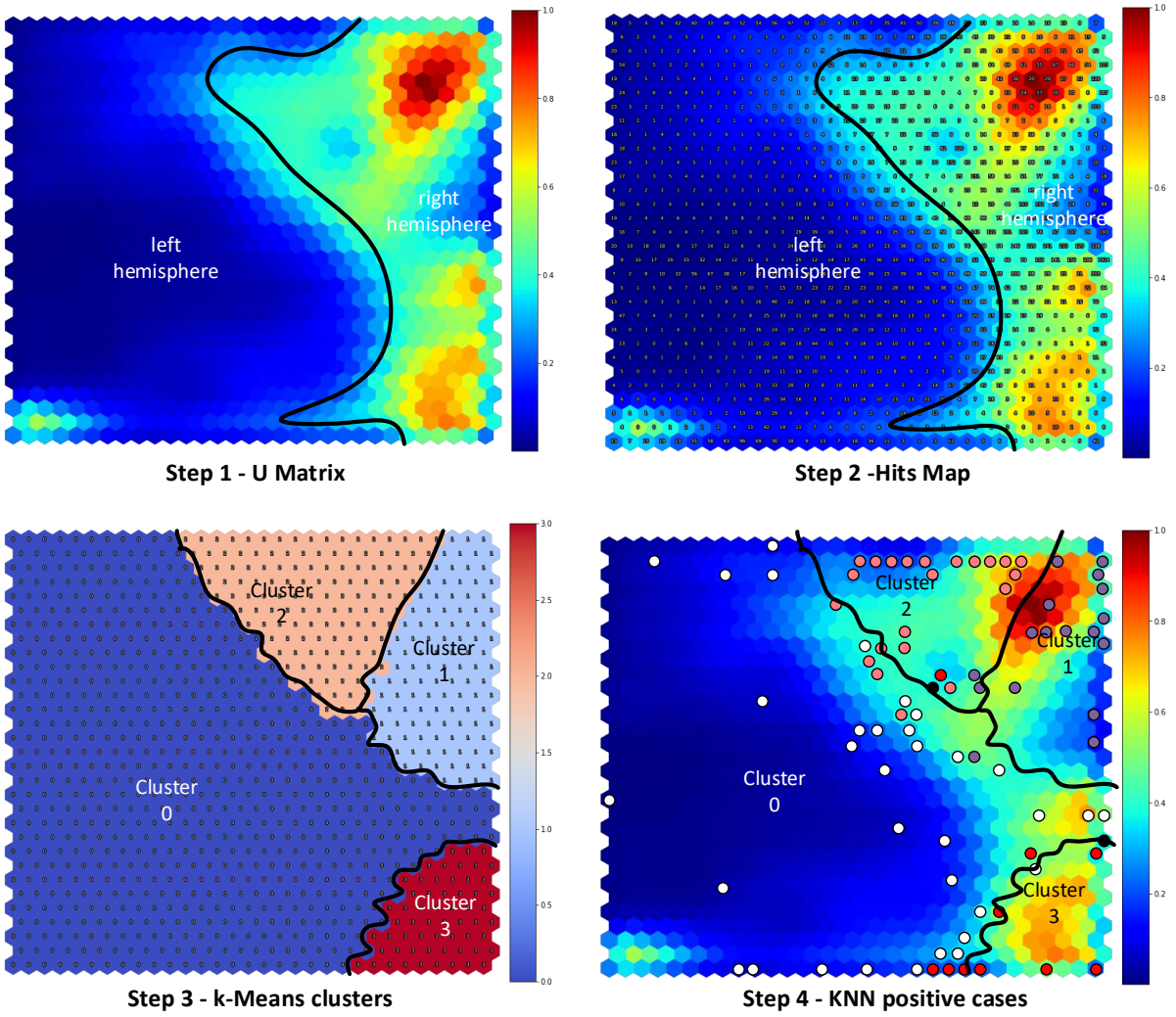

Figure 2 - Summary of the main steps followed to characterize VAT fraud

In Figure 2 above, the coloured dots (white – cluster 0; indigo – cluster 1; pink – cluster 2; red – cluster 3) show the location of neurons with positive cases isolated by the KNN.

Notice also some few black dots, representing neurons that superimpose onto one or more clusters of the k-Means, leading to an undetermined intermediate zone (Serrano-Cinca, 1998) of positive or non-positives cases. As exceptional cases, these neurons should be screened by a panel of MRA experts.

See also, in Figure 2, that the topography of the U Matrix clusters is not congruent with that generated by the k-Means, which is explained with the distortion caused by the topographic error (TE), in this case, $TE \approx 0.9$.

Thus, by crosschecking the KNN tagged-data with 27,049 taxpayers under the normal VAT regime, the result for each economic sector is (Table 3):

Table 3 - Consolidated Map of Frauds, Investigations and Suspicions of the Case Study

Economic Sector	Cluster 0			Cluster 1			Cluster 2			Cluster 3		
	F	**I**	**S**	**F**	**I**	**S**	**F**	**I**	**S**	**F**	**I**	**S**
A	1	0	6	1	0	2	0	0	1	0	0	0
B	**14**	**24**	**4**	**19**	**1**	**1**	**14**	**3**	**1**	**2**	**3**	**0**
C	4	1	2	0	0	0	4	0	0	10	0	1
D	2	1	1	2	0	0	2	0	0	1	0	0
E	1	0	2	0	0	0	0	0	1	1	0	0
F	**16**	**7**	**2**	**5**	**0**	**0**	**9**	**0**	**1**	**9**	**0**	**1**
G	2	1	0	0	0	0	0	0	0	1	0	0
Total:	40	**34**	**17**	27	**1**	**3**	29	**3**	**4**	24	**3**	**2**

From Table 3, it is clear that 41 taxpayers are marked as (F) and another 26 as (S), which proves hypotheses 1 and 2. A similar phenomenon happens, to a lesser extent, with partial results of the 2013-2017 and 2014-2018 audits. It is also noted that economic sectors A and E show a relatively high number of suspects (S) compared to the total number of frauds (F) reported by the MRA and a relatively high fraud detection rate in economic sector B, although accompanied by a significant number of taxpayers marked as investigations (I) and suspects (S).

On the other hand, thanks to four-step methodology, the initial audit sample is ultimately reduced to 9,088 taxpayers in 27,049, with 66% of positive cases confined to only 1,022 taxpayers (clusters 1, 2 and 3) belonging to the right hemisphere (Table 4):

Table 4 - Ratio of positive fraud cases by cluster

Cluster	Truly positive cases (F_n)	Frequency of taxpayers (T_n)	Ratio $\frac{F_n}{T_n} . 10^3$
0	40	7802	5.13
1	27	264	**102.27**
2	29	564	51.42
3	24	458	52.40

As the ratio of positive cases ($F_n, n = 0,1,2$ ...) is deducted from the frequency of taxpayers in each cluster ($T_n, n = 0,1,2$ ...), taxpayers grouped in clusters 1, 2 and 3 would be primarily targeted by audits and inspections, in descending order of the ratio $\frac{F_n}{T_n} . 10^3$, which can also help prioritize (Vanhoeyveld *et al.*, 2019) these actions in the field.

It is also implied that 0.44% precision achieved by the random method of the MRA (refer Table 1) is instantly increased to 6.22%, by simply applying four-step methodology to the Kohonen Map.

Finally, comparing the evaded VAT revenue (Mittal *et al.*, 2018) with that recovered by the Tax Authority of Mozambique, it turns out that exist[21], among the partial samples from 2013-2017; 2014-2018; 2017-2018 and the Case Study, highly significant ratios of VAT evasion (Table 5):

Table 5 - Comparison of evaded and recovered revenue

Sample	Ratio $\frac{R_s}{R_r}$	Remarks
2013-17 Audits	323.81	Dimensionality reduced with Fisher
2014-18 Audits	606.65	Dimensionality reduced with Decision Trees Variable Selectors (CART)
2017-18 Audits	0.00	Sample excluded, only 3 audits.
Case Study	122.09	Dimensionality reduced with Fisher

A trend that remains unchanged with a closer look of Case Study economic sectors (Table 6):

Table 6 - Comparison of evaded and recovered revenue by economic sector

Economic sector	Ratio $\frac{R_s}{R_r}$	Remarks
A	449.67	Cyclically neglected by audits and inspections
B	1304.56	Previously flagged as high fraud incidence
C	75.09	
D	59.17	
E	150.51	Cyclically neglected by audits and inspections
F	8.22	Previously flagged as medium-high fraud incidence
G	1444.41	

8. Conclusion

With the application of Kohonen Maps, the Case Study achieves 58.33% precision for fraud detection with a dimensionality reduction done with Fisher's Discriminant Analysis, confirming its effectiveness in predicting between normal and anomalous classes. This figure significantly exceeds fraud detection rates of the random method currently used by the MRA.

In addition, Kohonen Maps drastically reduce the initial sample taxpayers, with 66% of the spotted positive cases coming from three identified clusters, presenting itself as an instrument of choice for prioritizing audits and inspections.

Hypotheses 1 and 2 proved, when flagged out of the Case Study sample of 27,049 under the normal VAT regime, 67 new taxpayers matching very similar characteristics to those positive cases spotted with the weighting criteria of regular or unexpected audits by the MRA.

With a precision of 58.33%, hypothesis 3 is also proved, showing the revenue recovered by the MRA is far below the evaded one, a trend that does not change fine granularity analysis by economic sectors

[21] Applying the formula (8) to 58.33% precision of the Kohonen KNN anomaly detector algorithm.

of the Case Study, where becomes clear the disproportionality between the volume of audits, inspections and the payments due to VAT refunds. In short, a revenue recovery strategy that is poorly articulated with Public Finance expenditure.

As future studies, it is intended to analyse the impact of incorporating an Auto-Encoder in the model proposed here, to predict and characterize VAT fraud in southern Mozambique, namely, to deal with semi-structured and unstructured data generated by the interoperability that exists between the information systems of MRA and banking, public procurement, civil identification, commercial registration, and social security counterparts, among others.

9. References

BASTA, S. FASSETTI, M. GUARASCIO, M. MANCO, G. GIANNOTTI, PEDRESCHI, D. SPISANTI, L. PAPI, G. & PISANI, S. 2009. "High quality true-positive prediction for fiscal fraud detection". In 2009 IEEE International Conference on Data Mining Workshops.

BREUNIG, M.M. KRIEGEL, H. NG, R.T. & SANDER, J. 2000. “LOF: identifying density-based local outliers”. SIGMOD Rec. 29, 2 (June 2000), pp. 93–104.

DE ROUX, D. PÉREZ, B. MORENO, A. VILLAMIL, MDP & FIGUEROA, C. 2018. "Tax Fraud Detection for Under-Reporting Declarations Using an Unsupervised Machine Learning Approach". In KDD 2018, August 19-23, London, United Kingdom, pp. 215-222.

DUDA, R.O. HART, P.E. & STORK, D.G. 2000. "Pattern Classification. Second Edition”. Wiley, pp. 1-19.

FIX, EVELYN & HODGES, JOSEPH L. 1951.”Discriminatory Analysis. Nonparametric Discrimination: Consistency Properties”. USAF School of Aviation Medicine, Randolph Field, Texas.

GUARASCIO, M. 2010. "Data Mining Techniques for Fraud Detection". Tesi di Dottorato, Università della Calabria. Dipartimento di Elletronica, Informatica e Sistemistica. Dottorato di Ricerca in Ingegneria dei Sistemi e Informatica. Ciclo XXII.

HAJDÚCHOVÁ, I. SEDLIAČIKOVÁ, M. & VISZLAI, I. 2015. “Value-Added Tax Impact on the State Budget Expenditures and Incomes”. In Procedia Economics and Finance 34. Elsevier, pp. 676-681.

HUTTON, E. 2017. "The Revenue Administration - Gap Analysis Program: Model and Methodology for Value-Added Tax Gap Estimation". Fundo Monetário Internacional (FMI), Departamento de Assuntos Fiscais, pp. 3-25.

IMF. 2017. ”Digital Revolutions in Public Finance”. doi: https://doi.org/10.5089/9781484315224.071 (acedido em 25/05/2021)

JIHAL, H. TALHAOUI, M.A. DAIF, A. & AZZOUAZI, M. 2018. "Predictive Analytics as A Service on Moroccan Tax Evasion". In International Journal of Engineering & Technology, 7 (4.32) (2018), pp. 90-92.

JUPRI, M. & SARNO, R. 2020. "Data mining, fuzzy AHP and TOPSIS for optimizing taxpayer supervision". In Indonesian Journal of Electrical Engineering and Computer Science. Vol. 18, No. 1, pp. 75-87.

KIVILUOTO, K. & BERGIUS, P. 1998. "Maps for Analyzing Failures of Small and Medium-sized Enterprises". In G. Deboeck, & T. Kohonen (eds.), Visual Explorations in Finance with Self-Organizing Maps. Springer Finance. Springer, London. pp. 59-71.

KOHONEN T. 1998. “The SOM Methodology”. In G. Deboeck, & T. Kohonen (eds.), Visual Explorations in Finance with Self-Organizing Maps. Springer Finance. Springer, London. pp. 159-167.

KOHONEN, T. 1982. “Self-Organized Formation of Topologically Correct Feature Maps“. Biological Cybernetics. 43 (1): pp. 59–69.

KOHONEN, T. SCHROEDER, M.R. & HUANG, T.S. 2001. “Self-Organizing Maps (3rd. ed.)”. Springer-Verlag, Berlin, Heidelberg.

KREVER, R. 2008. "VAT in Africa". Pretoria University Law Press, pp. 71-80.

LA FERIA, R. & SCHOEMAN, A. 2019. "Addressing VAT Fraud in Developing Countries: The Tax Policy-Administration Symbiosis". In 47, Intertax, Issue 11, pp. 953-957; 961-962.

MACQUEEN, J. B. 1967. “Some Methods for classification and Analysis of Multivariate Observations“. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability. 1. University of California Press. pp. 281–297.

MANJATE, J.H. 2018. "Ineficiência da operacionalização do Sistema Fiscal Moçambicano". Tese de Mestrado, Instituto Superior de Contabilidade e Administração do Porto. Instituto Politécnico do Porto. pp. 54-57.

MATOS, T. MACEDO, J.A.F. & MONTEIRO, J.M. 2015. "An empirical method for discovering tax fraudsters: A real Case Study of Brazilian fiscal evasion". In Proceedings of the 19th International Database Engineering No. 38, Applications Symposium, in IDEAS '15, ACM, New York, NY, USA, 2014, pp. 41-48.

MITTAL, S. REICH, O. & MAHAJAN, A. 2018. "Who is bogus? Using one-sided labels to identify fraudulent firms from tax returns". In Proceedings of the 1st ACM SIGCAS Conference on Computing and Sustainable Societies, in: COMPASS'18, ACM, New York, NY, USA, 2018, pp. 24:1-24:11.

MWANZA, M. & PHIRI, J. 2016. "Fraud Detection on Bulk Tax Data Using Business Intelligence Data Mining Tool: A Case of Zambia Revenue Authority". In International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016, pp. 793-798.

PALMA, C.C. 2015."O Sistema de IVA em Moçambique: Adopção e Características Gerais". In Revista do Programa de Pós-Graduação em Direito da UFC. v. 35.1, pp. 379; 387-391.

PARPA II 2009. “PARPA II Review — The Tax System in Mozambique. Volume I”. United States Agency for International Development (USAID), pp. 29-31; 76-77.

PORTELA, A.P.R. 2014. "Fraude Fiscal em IVA". Trabalho do Curso de Pós-Graduação em Direito Fiscal. Universidade do Porto, pp. 7-28.

SCHÖLKOPF, B. WILLIAMSON, R. C. SMOLA, A.J. SHAWE-TAYLOR, & J. PLATT, J.C. 1999. "Support Vector Method for Novelty Detection". NIPS'99: Proceedings of the 12th International Conference on Neural Information Processing Systems. November 1999, pp. 582–588.

SERRANO-CINCA C. 1998. “Let Financial Data Speak for Themselves”. In G. Deboeck, & T. Kohonen (eds.), Visual Explorations in Finance with Self-Organizing Maps. Springer Finance. Springer, London. pp. 3-18.

SOTOMANE, C. 2014. "Factors Affecting the Use of Data Mining in Mozambique: Towards a framework to facilitate the use of data mining". Tese de Doutoramento In DSV Report Series No. 14-012. Stockholm Universitetsservice US AB, Stockholm. pp. 102-103.

STEINHAUS, H. 1957. “Sur la division des corps matériels en parties“. Bull. Acad. Polon. Sci. 4 (12): pp. 801–804

TIAN, J., AZARIAN, M. H., & PECHT, M. 2014. “Anomaly Detection Using Self-Organizing Maps-Based K-Nearest Neighbor Algorithm”. PHM Society European Conference, 2(1), pp.3-4.

TU, L.A. THAI, V.D. & HOAN, N. Q. 2016. "Improving Feature Map Quality of SOM Based on Adjusting the Neighborhood Function". International Journal of Computer Science and Information Security (IJCSIS), Vol. 14, No. 9, September 2016, pp 1-6.

ULTSCH,A.2003."U*-Matrix: a Tool to visualize Clusters in high dimensional Data". DataBionics Research Lab, Department of Computer Science University of Marburg, Tecnhical Report No.36, 2003.

UNICEF 2019."Fiscal Space Analysis ". UNICEF, Maputo, Mozambique. pp. 47-51.

VANHOEYVELD, J. MARTENS, D. & PEETERS, B 2019. "Value-Added tax fraud detection with scalable anomaly detection techniques". In Applied Soft Computing Journal 8640. Elsevier, pp. 1-20.

WENDEL, J. & BUTTENFIELD, B. 2010. "Formalizing Guidelines for Building Meaningful Self-Organizing Maps". In GIScience 2010: Sixth international conference on Geographic Information Science. Zurich, 14-17th September, 2010.

WU, R. OU, C.S. LIN, H. CHANG, S. & YEN, D.C. 2012. "Using data mining technique to enhance tax evasion detection performance". In Expert Systems with Applications 39. Elsevier, pp.8769–8777.