UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

FACULTAD DE INGENIERÍA

ANALYSIS OF INTERTEXTUAL DISTANCES USING MULTIDIMENSIONAL SCALING IN THE CONTEXT OF AUTHORSHIP ATTRIBUTION

ARTÍCULO ACADÉMICO

QUE PARA OBTENER EL TÍTULO DE:

INGENIERO EN COMPUTACIÓN

P R E S E N T A:

JULIÁN SOLÓRZANO SOTO

ASESORA:

DRA. FERNANDA LÓPEZ ESCOBEDO

(2016)

The final publication is available at Springer via http://dx.doi.org/10.1080/09296174.2016.1142324

Analysis of Intertextual Distances Using Multidimensional Scaling in the Context of Authorship Attribution*

Fernanda López-Escobedo[1], Julián Solorzano-Soto[2], Gerardo Sierra Martínez[3]

[1]Licenciatura en Ciencia Forense, Facultad de Medicina (UNAM), [2]Facultad de Ingeniería (UNAM), [3]Grupo de Ingeniería Lingüística, Instituto de Ingeniería (UNAM)

ABSTRACT

Four distance functions were evaluated in order to determine which better represents two types of style markers (named as static and dynamic) commonly used to authorship attribution tasks. Intertextual distances were analyzed from different authors and evaluate if the closest text to another was written by the same author. Classic multidimensional scaling was used to visualize intertextual distances because we consider that is a method that allows the judges to better understand and visualize the results.

The outcome of this paper is that selecting different distance functions considering the type of style marker improves the clustering of texts from the same author. While for static features we concluded that Canberra distance is recomendable, the dynamic features must depend on the style of each author.

Keywords

Authorship attribution, multidimensional scaling, stylometry, intertextual distance, forensic linguistics

*CORRESPONDING AUTHOR Address correspondence to Fernanda López-Escobedo, Licenciatura en Ciencia Forense, Facultad de Medicina (UNAM), Circuito Escolar s/n, Ciudad Universitaria, Delegación Coyoacán, México, D.F., C.P. 04510, Tel: +52 (55) 56 22 00 59, E-mail: flopeze@unam.mx

Address correspondence to Julián Solórzano-Soto, Facultad de Ingeniería (UNAM), Circuito Escolar s/n, Ciudad Universitaria, Delegación Coyoacán, México, D.F., C.P. 04510, Tel: +52 (55) 56 22 08 66, E-mail: jsolorzanos@hotmail.com

Address correspondence to Gerardo Sierra Martínez, Grupo de Ingeniería Lingüística, Instituto de Ingeniería (UNAM), Circuito Escolar s/n, Ciudad Universitaria, Delegación Coyoacán, México, D.F., C.P. 04510, Tel: +52 (55) 56 23 36 00 ext.8808, E-mail: GSierraM@ingen.unam.mx

1. INTRODUCTION

Forensic Linguistics has been defined as the interface between language and the law. Although this is a general definition, it encompasses all the areas studied in this field: language and law, language in the legal process, and language as evidence. Authorship attribution (AA), which is the interest of this work, is a task belonging to the latter of these three areas. This task refers to analyzing a written text in search of evidence that can lead to identifying its author.

According to Koppel, Schler, and Argamon (2009), there are at least three different scenarios in AA:

- When an unknown text (for which the author is unknown or disputed) must be attributed to an author among a set of different suspects.
- When an unknown text must be attributed or not to one author who is the only suspect.
- When there are no suspects and the task refers to providing information about the author's sociolinguistic characteristics (gender, age, etc.).

The first methods developed in authorship attribution studies were based on comparing the distribution curves of simple measures, such as word length frequency, across different texts. The aim of these methods was to define a measure to differentiate between a set of authors. Koppel et al. (2009) call this the 'unitary invariant' approach. Soon, these methods were proved to be unreliable and the multivariate approach was introduced. This new approach involved the comparison of several measures at the same time. Mosteller and Wallace's work (1964) is commonly cited as the beginning of this new wave of modern AA methods, in which a wide variety of data is taken into account. In order to analyze such an amount of information (known as high-dimensional data) several mathematical and statistical techniques have been proposed. These techniques often give researchers a visual representation of the data, which is very helpful in revealing its internal structure, and in this way they may discover patterns that otherwise would have been missed. Burrows (1987), who used eigen analysis, is often cited as the first work that used a visualization technique to analyze writing style. In more recent examples, the 'Writeprints' method developed by Abbasi and Chen (2008) creates, for a given text, a visual pattern which represents the writing style of the author. A pattern can also be obtained from an unknown text and then compared to previously generated patterns obtained from texts with known authors. In this way, AA becomes similar to matching a suspect's fingerprints to those in a database. Another experiment in visual textual analysis was performed by Keim and Oelke (2007). In their experiment, they analyzed books written by two distinct authors. Every chapter of each book was assigned a color depending on certain stylistic measurements. When the analysis is visualized, the author of the book can be guessed by the overall color of the resulting images. Inspired by these approaches, in this study we experiment with a well-known method for information visualization called multidimensional scaling (MDS).

MDS allows researchers to visualize distances between a set of objects. For example, applying MDS to a matrix containing the distances in kilometers between every city in Europe will result in a set of two dimensional points. If these points are plotted, an image very similar to an actual map of Europe will be generated. In the AA task, the matrix has to

contain the distances between the text documents, known as intertextual distance. Previous experiments relating to intertextual distance include the one done by Labbé (2007), who used a tree classification graph to visualize the resulting relationships between the texts. Another example, is provided by Merriam (2003), who used MDS itself to compare plays by Shakespeare and Middleton.

We think that the concept of distances between objects is a very intuitive one. A linguist called to analyze evidence consisting of a disputed text in a legal case, is commited to explaining how he or she arrives to a certain conclusion. Usually the judge and jury do not have expertise in forensic linguistics, and therefore a visual representation of the results is very helpful. As explained, MDS is able to represent the distances between a set of objects (in this case, the text documents) in a low-dimensional space (in this case, a two dimensional space). We consider that the intuitive nature of the methodology is an important factor when explaining it to non-experts. The intuition is that texts which are similar in writing style, and thus probably written by the same author, are 'close' to one another and 'far' from texts written by another author.

In order to calculate the intertextual distance, each text must be converted into a point in a vector space. Each point consists of a series of numbers obtained from stylistic measurements. This style quantification process is called stylometry and it is discussed in Section 2. Then, a distance function must be chosen to assign a distance value between the points. This measure will express how 'close' or 'far' the texts are from each other. In this study, we evaluated different standard distance functions (described in Section 3) in an empirical resource described in Section 4. Also, we relate the type of stylometric measure with the type of distance function, which has an effect on the results as can be seen in Section 5. Finally, we present the conclusions in Section 6.

2. STYLOMETRY

In broad terms, stylometry is understood as the study of style quantification (Golcher, 2007); that is, the extraction of the most appropriate features which can provide quantitative information about an author's style (Stamatatos, 2006). Stylometry relies on the assumption that it is possible to consistently identify an author's style by examining his or her linguistic choices (Guillen-Nieto, 2008), which are known as stylometric features (Madigan, Genkin, Lewis, Argamon, Fradkin, & Ye, 2005), style markers (Peng, 2003), or simply as measures. In this work, the term style marker is used to refer to a category or set of certain individual style features. For example, an individual feature is the frequency of occurrence of commas; another one is the frequency of occurrence of semicolons, and so on. All these features can be classified as belonging to a style marker named 'frequency of punctuation marks'.

One of the main challenges in stylometry is to decide which features will be used. Rudman (1997) states that more than one thousand features have already been identified with no consensus regarding which are the best. He argues that one should strive to find a complete set of features: a 'mapping of style' much like the mapping of genes in DNA. Novel features are constantly being proposed, such as Measure S (Golcher, 2007), flexible patterns and k-signatures (Schwartz, Tsr, Rappoport, & Koppel, 2013). Furthermore, there are features that are specific to certain domains. For example, for online authorship attribution (in which the

texts are emails, chat logs, forum posts), certain structural elements can be used as style markers such as HTML tags (De Vel, Anderson, Corney, & Mohay, 2001).

It is common for researchers to present a classification of style markers in their experiments or surveys. Table 1 exemplifies the difference in the taxonomies used by different authors.

This work's objective is to test different distance functions depending on the linguistic data and evaluate the usefulness of MDS as a technique to assist in authorship attribution studies, and not to propose an approach to select the set of features which better characterizes the style of an author. Therefore, the analyzed style markers were selected from the most common style markers used in previous studies as shown in Table 1. There does not necessarily have to be a very large number of them. Experiments such as Ruseti's (2012) have already shown that a reduced feature set still yields good results. In addition, our selection of style markers takes into account the observations that other researchers have made. For example, Golcher (2007) considers syntactic features to be underrepresented in most studies. Therefore, two forms of POS tags were analyzed as an attempt to capture more varied syntactic information. Table 2 shows the 15 style markers analyzed in this study.

We have placed the style markers into two categories, in accordance with the nature of the features each one encompasses. Following Abbasi and Chen (2008), we distinguish between static and dynamic style markers. Features of static style markers are context-free and well-defined categories, such as the set of punctuation marks. On the other hand, features of dynamic style markers, such as n-grams, are context-dependent features and have infinite potential feature spaces. We looked more into this distinction and concluded that an important difference is that the latter tend to produce vectors containing a lot of zeros because it is less probable that all texts present them. The density of the features matrix (ratio of non-zero entries to total entries) of each style marker was analyzed to confirm this difference. For purposes of this study, the style markers with a density greater than 0.5 were considered as static and those with a density lower than 0.5 were considered as dynamic. Table 2 shows the density of each style marker. It is important to notice that the lexical features marker presents a density of 1, since all its features have values for all texts, while content word trigrams presents the lowest density, as most trigrams are unique to each text.

We employed the open source language analysis tool Freeling (Padró & Stanilovsky, 2012) to generate the POS tags, which use the EAGLES tagset. Two different forms of POS tag style markers were defined. One was using the full tag, which contained detailed information of each word such as gender, number and tense. The other one was using only the first letter of the tag, which corresponds to the most general grammatical category. Table 3 shows an example of both POS tag style markers applied to a sentence. Bigrams and trigrams were also analyzed with both forms of POS tags.

3. MULTIDIMENSIONAL SCALING AND DISTANCE FUNCTIONS

One of the purposes of MDS, described by Borg and Groenen (2005), is to represent dissimilarity data as distances in a low-dimensional space in order to make this data accessible to visual inspection and exploration. The more dissimilar the objects are to one another, the greater the distance value must be between them. For the AA task, the objects

being analyzed are the text documents. The dissimilarity between any two texts is directly proportional to the difference between the measurements of their stylometric features. The dissimilarity is quantified using a distance function.

There are different distance functions available for quantitative data. One of the most common, perhaps because it represents our physical concept of distance, is the Euclidean distance. Let X be the set of all the vectorized text documents. X is composed of points in m dimensions, each point representing a single text document, and each dimension being the relative frequency of the occurrence of feature a in that document. Then, the Euclidean distance d_{ij} between points x_i and x_j is given by

$$d_{ij}(X) = \left(\sum_{a=1}^{m} \left(x_{ia} - x_{ja}\right)^2\right)^{\frac{1}{2}} \tag{3.1}$$

The Euclidean distance, along with the city-block distance (also called Manhattan distance), is a special case of the Minkowsky distance represented by the following formula:

$$d_{ij}(X) = \left(\sum_{a=1}^{m} \left| x_{ia} - x_{ja} \right|^p\right)^{\frac{1}{p}}, p \geq 1 \tag{3.2}$$

For *p*=2 we obtain (3.1) and for *p*=1 we obtain the city-block distance shown in (3.3):

$$d_{ij}(X) = \sum_{a=1}^{m} \left|x_{ia} - x_{ja}\right| \tag{3.3}$$

The Canberra distance is another measure for quantitative data represented by (3.4):

$$d_{ij}(X) = \sum_{\mathrm{a=1}}^{\mathrm{m}} \frac{\left|x_{ia} - x_{ja}\right|}{\left|x_{ia} + x_{ja}\right|} \tag{3.4}$$

where distance between points *i* and *j* is equal to the sum of the absolute values of the Difference between the feature frequencies divided by their sum.

Besides these distance functions for quantitative data, similarity coefficients for binary data such as the Jaccard distance also exist:

$$s_{ij} = \frac{b+c}{a+b+c}, 0 \leq s_{ij} \leq 1 \tag{3.5}$$

where *a* is the number of features which are present in both texts, and *b* and *c* are the number of those which appear only in one of the two texts. Note that in this kind of distance, it is not the frequency of the feature which is measured but rather only whether that feature occurs in the text or not. That is, the value of each dimension is either 1 or 0.

There is no agreement about what kind of distance function is better for analyzing the type of data in a given study. It is true that there are different types of functions depending on whether the data is quantitative or binary. Yet, to decide, for example, if the Euclidean distance is better than the City-Block distance for a particular case is not straight-forward. Therefore, one of the purposes of this work is to analyze data for authorship attribution with different distance functions in order to conclude which is better for different stylometric features.

A binary distance such as the Jaccard distance was expected to better represent dynamic style markers because the main source of information is whether the feature is present or not as opposed to its actual frequency. This is adequate considering that it is not very probable that dynamic features will be present in all texts because they are context-dependent features. Regarding static type markers, quantitative distances were expected to yield better results since they are a closed set of features and thus it is highly probable that they will be present in all texts.

We decided to test four different distance functions to generate the distance matrix for the MDS: the Euclidean, the City-Block, the Canberra, and the Jaccard distance. The first two combine dimensional differences directly; therefore if these dimensions are values measured on different scales it is desirable to standardize the values and avoid different variances between them (Borg & Groenen, 2005). The data analyzed in this work represent the relative frequency of occurrence of different style markers among a set of texts, therefore the values are standardized. However, some features are present in one text but not in another one. For this reason, the Canberra distance, which has some provision for controlling the dispersion either for each variable separately or for all variables simultaneously (Borg & Groenen, 2005), was also tested.

Once the distance has been generated, the next step is creating the metric MDS model. This class of MDS model preserves the data linearity in the distances (Borg, Groenen, & Mair, 2013) and is represented by:

$$p_{ij} \rightarrow a + b \cdot p_{ij} = d_{ij}(\mathbf{X}) \tag{3.6}$$

where the distance is the result of a linear transformation of the distances between two objects *i* and *j*. It is assumed that no information of the data is lost if multiplied by an arbitrary constant *b* or if a constant *a* is added to each data value.

In order to generate the metric MDS model, as well as the distance matrix, we used R (R Core Team, 2014), a free software environment for statistical computing.

4. CORPUS

The empirical resource created for this work is composed of nine texts by six different authors. They were all born in Mexico except for José de la Colina, who was born in Spain in 1934 but has lived in Mexico since 1940. All the authors published over a span of 23 years, from 1990 to 2013. The texts from three different genres: short story, article, and essay are included. The textual data length is between 408 and 9632 words, with an average of 1824 words. Table 4 shows the structure of the corpus.

The corpus' design allows to test if MDS is a suitable technique for AA independent of text genre. Also, it is possible to evaluate whether or not it is suitable in spite of short texts being present. As Stamatatos (2009) pointed out, the textual data length is an important methodological issue in authorship attribution. Traditionally AA studies were done on long texts such as novels. Nevertheless, in realistic scenarios, texts tend to be much shorter. Therefore, in recent years more research has focused on the attribution of short segments of text such as in Metzler, Dumais, and Meek (2007) and in Schwartz et al. (2013).

The corpus was codified in UTF-8 and lemmatized. All style markers were set to use the lemmatized tokens except for the case of the features marked as 'lexical features' and all of those involving full POS tags. The relative frequency of the chosen style markers' features were extracted for each text. These data was used to generate the distance matrix in order to perform the MDS. In the following section we present the results obtained in this study.

5. RESULTS AND DISCUSSION

When a particular set of style markers is used to discriminate between different authors, we do not expect the same set of style markers to have the same discriminant potential to characterize all authors. As (AUTHOR) mentioned, often a set of style markers will be more effective for characterizing a particular author but will be less effective for other authors. That is, the optimal set of features is dependent on each particular author's style. Additionally, as other studies have shown (Grieve, 2007; Madigan et al., 2005), when the number of authors is increased, the style markers reduce their discriminant potential. In accordance with these two ideas, when all the style markers are used to generate the MDS it is not possible to visualize differences between all the authors in the corpus (see Figure 1). Nevertheless, texts from the same author are not scattered all over the graph. For example, in both graphs all the texts by José de la Colina (JC) are plotted at the top.

Considering these issues, when more than two authors are analyzed, the problem is treated as a binary classification where one class corresponds to the question candidate and the other corresponds to texts from other authors known as impostors. This is in order to answer the question: are the questioned texts written by the questioned author? where the answer is yes or no, rather than attributing the text to one of the authors in the corpus. This corresponds to the open class scenario of the authorship attribution task. When texts are plotted using MDS, the region in which the disputed text appears gives elements to the linguist to draw a conclusion about the authorship of the text. One of regions must correspond at the suspect and the other at the impostor or impostors, depending on the number of authors in this class. In this work, the class containing the impostors can vary in size from one to five authors (from 9 to 45 texts).

Additionally, instead of performing the experiments using just all features simultaneously, it was performed using either only static or only dynamic features at once.

In order to facilitate the visualization of differences in results depending on the distances and the features analyzed plots with two authors are presented. The last section includes the results obtained when the number of the authors in the impostors class is increased.

Static Features

As mentioned before, quantitative distances were expected to yield better results with static features since they are a closed set of features and thus it is highly probable that they will be present in all texts.

Figure 2 displays one of the MDS plots obtained to represent distances between two authors: Alberto Chimal (AC) versus Enrique Serna (ES). The MDS was generated with static features using the Euclidean and the Jaccard distances. As expected, in plot generated with Euclidean distance (Figure (2a)) it is possible to draw a straight line to divide the graph into two regions, each one containing the texts by one autor. This is not possible in plot generated with Jaccard distance (Figure (2b))

The Manhattan distance was also evaluated as a function to represent static features. Results shows that this distance yield very similar results as Euclidean distance. Nevertheless, in some cases Manhattan results better to distinguish two regions in the graph as shows in Figure 3. Using Euclidean distance (Figure (3a)) text JC6 is plotted in the region of EP texts and text EP8 in region of JC. When we used Manhattan distance (Figure (3b)) those texts are closed to each other but it is posible to visualize two regions corresponding to texts of JC at the top and texts from EP at the bottom.

When Canberra distance is used to analyze static features the results show that this distance improves the diferentiation between regions of texts by each autor. For example, using Canberra distance to analyze the same pair of authors represented in Figure 3 it is shown that Canberra distance (Figure 4) improves the visualization of two different regions containing texts from each autor. This occurs in almost all of the cases.

Dynamic Features

Dynamic style markers were expected to be better handled by a binary distance such as the Jaccard distance. For example Figure 5 display two MDS plots for texts from MB versus AC and texts from MB versus EP using dynamic features. Figure 5 shows that the use of a dissimilarity measure for binary data when dynamic features are analyzed improves visualization. Texts for each author are clearly separated in this figure. It is possible to draw a straight line dividing Figure (5a) into a top and bottom section. All the texts by MB are plotted in the top and all the texts by AC are in the bottom section. In Figure (5b) a diagonal line dividing regions corresponding to texts by MB and texts by EP can be drawn. It is also worth noting that the three story texts by EP (EP1, EP2, and EP3) appear considerably farther away from the other EP texts, yet they do not overlap with those by MB. This suggests that it is possible to have genre-independent AA in some cases, given that the author regions are still clearly delimited.

Figure 5 shows that dynamic style markers are better represented with a Jaccard distance, however this not occurs in all cases. For example, in Figure 6 a MDS is displayed for texts from AM and JC using dynamic variables and employing the Manhattan (Figure (6a)) and the Jaccard distance (Figure (6b)). It is possible to distinguish two different regions corresponding to texts by each autor in both graphs. Similar results occurs in other pairs of authors, where dynamic features are well represented sometimes with a binary distance and sometimes with a quantitative distance. The only constant is that dynamic features are well represented in all cases using Canberra distance as it occurs with static features.

In order to test the capacity of the Canberra distance to represent all variables, both static and dynamic features we analyzed at the same time. Figure 7 shows that the Canberra distance succeeds in creating two regions in the graph. This makes it seem as though there is no need to distinguish between static and dynamic variables when Canberra distance is used.

This result was compared to the Canberra performance when analyzing dynamic features. Figure 8 shows that using the Canberra distance with only dynamic features also gives a very similar result. This suggests that when using the Canberra distance, the information that static features contribute to the visualization is little to none. In order to account for this loss of information, the static features have to be analyzed separately.

Increasing Number of Authors

Experiments considering all possible combinations of style markers and class schemes were performed. In each experiment we vary: a) the size of the impostor class and b) the author of the non-impostor class. Since the size of the impostor class can be 1, 2, 3, 4 or 5 and the author of the non-impostor class (the questioned author) can be one of the six different authors in our corpus, the total number of class schemes is 30. Furthermore there are a number of combinations to consider regarding the authors present in the impostor class. For the impostor class of size 1 there are 5 possible cases (5 different autors), for the class of size 2, there are 10 possible combinations and so on. For each class scheme the average of all combinations is considered.

The style markers to be used in each experiment can be a) static b) dynamic and c) both, and the distance to be used can be one of the four already mentioned: a) euclidean, b) manhattan, c) Jaccard and d) Canberra. This gives us a total of 12 style marker-distance combinations, when which multiplied by the 30 class schemes gives us a total of 360 experiments.

Given the large amount of plots that would to be analyzed, an automatic method of measuring how well the questioned author is separated from the impostors was used. For this, a score was calculated by counting the number of times that a text from the questioned author had as its closest text, one from that same author and divided it by 9, which is the number of documents of each author (the score was then averaged for all combinations of authors in the impostor class). So the perfect score is achieved when all texts from the questioned author are next to each other.

The average of all 360 experiments shows that Canberra distance is more effective with static features compared to Euclidean and Manhattan as displayed in Table 5. The average scores obtained with static features using Canberra results in an 83% of texts correctly classified followed by Manhattan with the 78% of correct classification.

Regarding the results of dynamic features, they are not as stable as static features and there are more variability on the results depending on the author and even on the distance metric. For example, when AC is the questioned author, dynamic features are better with Manhattan than with binary distance as shown in Table 6. However, when JC is the questioned author, dynamic features are not good regardless of the distance function. As noted in the previous section the Canberra is the distance that in almost all cases performs the best. In four of the six authors (ES, EP, MB and JC) the best results include Canberra distance. In AC the result obtained with Canberra distance is similar to the best result obtained with Manhattan, with only 3% of difference.

This keeps on reinforcing the idea that each author will be better characterized by a different set of features, while also suggesting that this is true more for dynamic features than for static features, which consistently have good results using Canberra distance for all authors.

When the number of authors in the impostor class is increased, the percentage of the number of times that a text from the questioned author has as its closest text one from that

same author is reduced as shown in Table 7. When the number of authors in the impostor class is only one, in 94% of the cases the closest text was one from the questioned autor. This percentage is 7% less when the number of authors in the impostor class is five. Each score in this table is an average of the best scores obtained for each author using the four distance functions with static markers and the four distance functions with dynamic style markers, i.e. for each size of the impostor class, the average of the best of the authors' eight scores. Although the average is reduced when the number of impostors are increased, a 87% is still a good result if 7 of the 9 texts has as its closest text one from that same author.

6. CONCLUSION

The objective of this paper was to evaluate different distance functions depending on the type of style marker. The results show that differentiating between static and dynamic features in order to apply a particular distance function reveals important clues in the author style. It is recommended to use Canberra distances when the style markers consist of a closed set of features (static features), whereas, when style markers are dynamic, in depends more on the questioned author. In general using a Canberra distance is better than using other distances, but this can change depending on the style of the author. There can be authors that do not use dynamic features that outstand from the rest, such as José de la Colina in our corpus (see Table 6).

Distinguishing between static and dynamic style markers is fundamental because using the two types at the same time will result in a loss of information. Taking both static and dynamic style markers and applying the Canberra distance causes the static style markers to be, for the most part, ignored, whereas applying the Euclidean or Manhattan distances reduces the discriminatory potential of both.

Regarding the effectiveness of the visualization technique when explaining distances between texts to non-experts, the results show that it is possible to visualize different regions containing texts by the same author. If the disputed text is plotted inside the region of the suspect, the linguist will have more elements to conclude that the features of this text are very similar to the ones of the suspect's texts and were therefore probably written by the same author. Otherwise, if the disputed text is plotted far away from the suspect's regions, the linguist could conclude that the style of the two authors is not similar to the one in the unknown text and therefore attributing authorship will not be possible.

These experiments confirm that using distances generated from style features is a good approach to classifying texts by author. The distance to be used depends on the type of features. While for static features we have concluded that Canberra distance is recomendable, the question remains for the case of dynamic features and whether it holds for any author. Thus further work can focus on one hand on the automatic learning of the distance function (techniques known as distance metric learning) and on the other on using proper classifying algorithms to take advantage of this and be able to perform authorship attribution with unknown texts in an automatic fashion. For example, an ensemble classifier could be used, in which each classifier specializes in a different kind of style marker and uses the most appropriate distance function.

ACKNOWLEDGEMENTS

This work was supported by the Mexico's National Council of Science and Technology (Conacyt) under Grant number 178248 and Project UNAM-DGAPA-PAPIIT under Grant number IN400312.

REFERENCES

Abbasi, A., & Chen, H. (2008). Writeprints: A stylometric approach to identity level identification and similarity detection in cyberspace. *ACM Transactions on Information Systems (TOIS), 26*(2), 7:1-7:29.

Borg, I., & Groenen, P.J. (2005). *Modern Multidimensional Scaling*, 2nd edition. New York: Springer.

Borg, I., Groenen, P. J., & Mair, P. (2013). *Applied Multidimensional Scaling*. Springer Science & Business Media.

Burrows, J. F. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. Literary and linguistic Computing, 2(2), 61-70.

Golcher, F. (2007). A new text statistical measure and its application to stylometry. In Matthew Davies, Paul Rayson, Susan Hunston & Pernilla Danielsson (eds.), *Proceedings of the Corpus Linguistics Conference (CL2007).*

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing, 22*(3), 251 – 270.

Guillén Nieto, V., Vargas Sierra, C., Pardiño Juan, M., Martínez Barco, P., & Suárez Cueto, A. (2008). Exploring state-of-the-art software for forensic authorship identification. *IJES, International Journal of English Studies*. *8*(1), 1-28.

Keim, D. A., & Oelke, D. (2007). Literature fingerprinting: A new method for visual literary analysis. *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE VAST 2007*, 115–122.

Koppel, M., Schler, J., & Argamon, S. (2009). Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology, 60*(1), 9 – 26.

Koppel, M., & Schler, J. (2003). Exploiting stylistic idiosyncrasies for authorship attribution. *Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis,* 69-72.

Labbé, D. (2007). Experiments on authorship attribution by intertextual distance in English. *Journal of Quantitative Linguistics, 14*(1), 33 – 80.

Lluís Padró and Evgeny Stanilovsky. (2012) FreeLing 3.0: Towards Wider Multilinguality Proceedings of the Language Resources and Evaluation Conference (LREC 2012) ELRA. Istanbul, Turkey. May, 2012.

Madigan, D., Genkin, A., Lewis, D. D., Argamon, S., Fradkin, D., & Ye, L. (2005, June). Author Identification on the Large Scale. Paper presented at the Annual Meeting of the Classification Society of North America, St. Louis, MO.

Merriam, T. (2003). An application of authorship attribution by intertextual distance in English. *Corpus, 2,* 167-182.

Metzler, D., Dumais, S., & Meek, C. (2007). Similarity measures for short segments of text. *Proceedings of the 29th European Conference on Information Retrieval*, 16–27.

Mosteller, F., & Wallace, D.L. (1964). Inference and disputed authorship: The Federalist. Reading. MA: Addison-Wesley.

Peng, F., Schuurmans, D., Wang, S., & Keselj, V. (2003, April). Language independent authorship attribution using character level language models. *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics, 1*, 267-274.

Rudman, J. (1997). The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities, 31*(4), 351 – 365.

Schwartz, R., Tsr, O., Rappoport, A., & Koppel, M. (2013). Authorship attribution of micro messages. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1880 – 1891.

Stamatatos, E. (2006). Authorship attribution based on feature set subspacing ensembles. *International Journal on Artificial Intelligence Tools, 15*(5), 823 – 838.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology, 60*(3), 538 – 556.

R Core Team (2014). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ruseti, S., & Rebedea, T. (2012). Authorship Identification Using a Reduced Set of Linguistic Features. Notebook for PAN at CLEF 2012, 2012.

De Vel, O., Anderson, A., Corney, M., & Mohay, G. 2001). Mining e-mail content for author identification forensics. *ACM Sigmod Record, 30*(4), 55 – 64.

Group	Features		
	Abbasi & Chen, 2008	*Koppel & Schler, 2003*	*Statamatos, 2009*
Lexical	Word-level, character-level, character n-grams, digit n-grams, word length distribution, vocabulary richness, special characters	Top function words	Word length, sentence length, vocabulary richness, word frequencies, word n-grams, errors
Character			Letters, digits, character n-grams, compression methods
Syntactic	Function words, punctuation, POS tags n-grams		POS tags, chunks, sentence and phrase structure, rewrite rules, frequencies, errors
Part-of-speech tags		POS tags, POS tag bigrams that appear at least 3 times	
Semantic			Synonyms, semantic dependencies
Structural	Message-level, paragraph-level, technical structure		
Content	Words, word bigrams, and word trigrams		
Idiosyncratic	Misspelled words	Syntactic (run-on sentences, etc.), formatting (ALL CAPS), spelling	
Application specific			Functional, structural, content-specific, language-specific

Table 1. Comparison of different classifications of style markers. For Abbasi and Chen, function words are syntactic features, whereas for Koppel and Schler they are lexical features.

Style markers	Density	Type
Punctuation	0.531	Static
Lexical features[1]	1	Static
Single-position word profile[2]	0.640	Static
Function words	0.510	Static
Function word bigrams	0.094	Dynamic
Function word trigrams	0.032	Dynamic
Content words	0.043	Dynamic
Content word bigrams	0.020	Dynamic
Content word trigrams	0.018	Dynamic
POS tags (full tag)	0.377	Dynamic
POS tag bigrams (full tag)	0.130	Dynamic
POS tag trigrams (full tag)	0.058	Dynamic
POS tags (reduced tag)	0.880	Static
POS tag bigrams (reduced tag)	0.638	Static
POS tag trigrams (reduced tag)	0.263	Dynamic

[1] Includes word and sentence length distribution, type token ratio and hapax legomena.
[2] Refers to the relative frequency of the POS tags being in the first position of a sentence.

Table 2. Style markers analyzed in this study.

	La luna es blanca			
	La	*luna*	*es*	*blanca*
POS tags (full tag)	DA0FS0	NCFS000	VSIP3S0	AQ0FS0
POS tags (reduced tag)	D	N	V	A

Table 3. POS tags example for the sentence “La luna es blanca” (*The moon is white*).

Corpus Structure									
	Story			Article			Essay		
	Title	Code	Tokens	Title	Code	Tokens	Title	Code	Tokens
Angeles Mastreta (AM)	La tía Chila	AM1	840	Contar los días	AM4	933	Guiso feminista	AM7	1778
	Mujeres de grandes ojos	AM2	621	Ver más allá	AM5	1344	La mujer es un misterio	AM8	2552
	La tía Mónica	AM3	617	Con el cuerpo en el aire	AM6	1629	Soñar una novela	AM9	2362
Alberto Chimal (AC)	El juego más antiguo	AC1	855	Lo fantástico en México: la vida al margen	AC4	1160	La idea de México	AC7	962
	La pasión según la sombra	AC2	6457	JLB y la CF	AC5	2502	La ciudad invisible	AC8	1795
	Mogo	AC3	6532	El carnaval de Ray Bradbury	AC6	1109	Manifiesto de un cuento mutante	AC9	1793
Eduardo Antonio Parra (EP)	Encuentro nocturno	EP1	2892	Vergüenza	EP4	1139	La libertad y la locura	EP7	1908
	La condena	EP2	4160	De filósofos y tiranos	EP5	960	El mono desnudo	EP8	1925
	Después del agüacero	EP3	1771	Conversaciones para encontrar el olvido	EP6	1891	La vida continúa	EP9	1031
Enrique Serna (ES)	Vanagloria	ES1	9632	La rebelión asexual	ES4	1560	El naco en el país de las castas	ES7	1477
	Drama de honor	ES2	4442	Patriotismo inducido	ES5	1080	La deshumanización del antro	ES8	1162
	La incondicional	ES3	4082	Finísimas personas	ES6	1018	Inducción a la santidad	ES9	975
José de la Colina (JC)	La princesa del café de chinos	JC1	1723	¿Una de las 7 ciudades maravilla?	JC4	427	Siempre tendremos Casablanca	JC7	928
	Muchacha del vestido color mamey	JC2	911	Al fin la FIL del Zócalo	JC5	408	Cuando el cine abolió la muerte	JC8	576
	La aventura del señor Loredo	JC3	1080	El emperador, el rayo y el ala de la Victoria	JC6	799	La invención de Drácula	JC9	943
Mario Bellatín (MB)	Biografía fantasma	MB1	1343	La mujer del analista	MB4	1141	Snapshots	MB7	1400
	El Cardenal y el Tapacaminos: un vacío poblado de nada	MB2	1300	La eternidad de la condena, de Horacio Ortiz	MB5	802	Manual para los devotos de Sergio Pitol	MB8	1139
	Un cuento de pingüinos	MB3	1127	Madre e hijo	MB6	1905	Giradores en torno a mi tumba	MB9	1573

Table 4. Structure of the corpus.

	Euclidean	**Manhattan**	**Jaccard**	**Canberra**
Average	0.6833	0.7879	0.5346	0.83878

Table 5. Score of static features using the four different distances.

	AC	AM	ES	EP	MB	JC
Euclidean	0.7491	0.6559	0.9426	0.9426	0.4193	0.2401
Manhattan	0.7706	0.1971	1	0.7706	0.1326	0.0788
Jaccard	0.7132	0.2258	1	0.9426	0.5698	0.3799
Canberra	0.7419	0.4695	1	0.9426	0.7706	0.4587

Table 6. Score of dynamic features in each author using the four different distances.

Number of authors in the impostor class	Average of the best results of all authors
1	0.9379928
2	0.9166667
3	0.8925926
4	0.8777778
5	0.8703704

Table 7. Average of the best results of all authors increasing the number of authors in the impostor class.

FIGURE CAPTIONS:

Fig. 1. MDS with all the features using the Euclidean (1a) and the Jaccard distances (1b) for all authors.

Fig. 2. MDS with static features using the Euclidean (2a) and the the Jaccard distances (2b) for one pair of authors: AC and ES.

Fig. 3. MDS with static features using the Euclidean (3a) and the Manhattan distances (3b) for one pair of authors: EP and JC.

Fig. 4. MDS with static features using the Jaccard distance for one pair of authors: EP and JC.

Fig. 5. MDS with dynamic features using the Jaccard distance for two pairs of authors: AC and MB (5a), EP and MB (5b).

Fig. 6. MDS with dynamic features using the Manhattan (6a) and the Jaccard distances (6b) for one pair of authors: AM and JC.

Fig. 7. MDS with all the features using the Canberra distance for one pair of authors: JC and EP.

Fig. 8. MDS with dynamic features using the Canberra distance for one pair of authors: JC and EP.

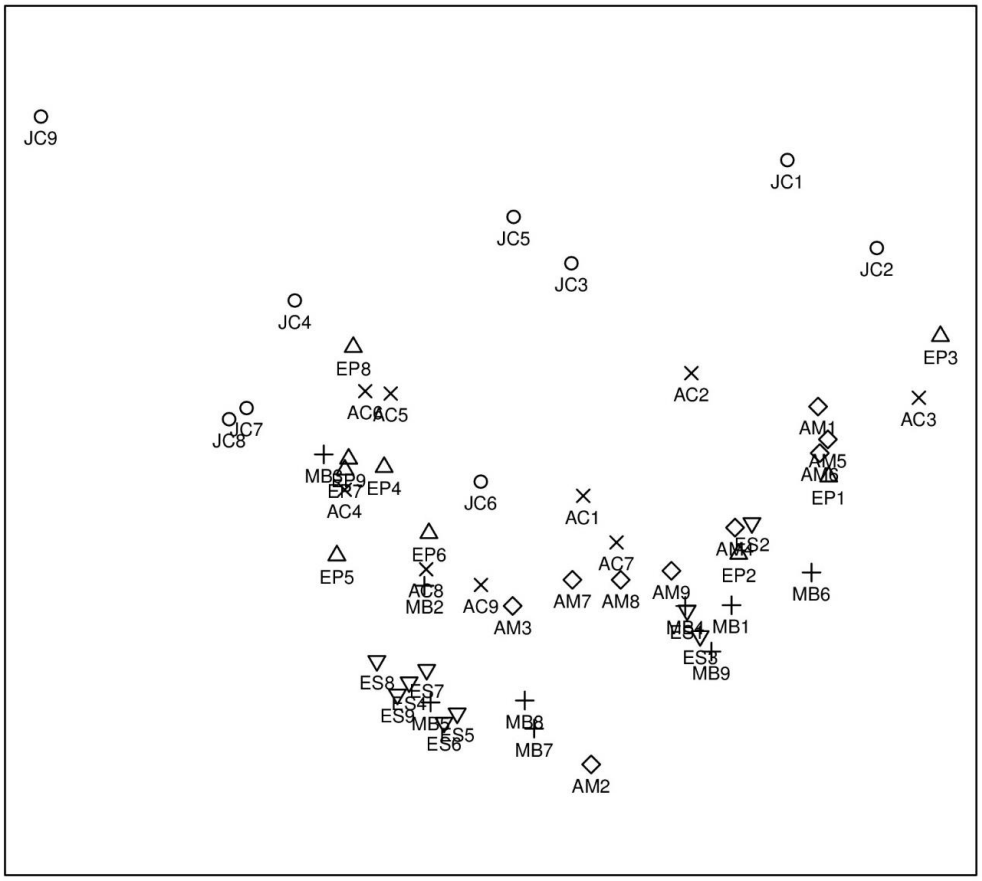

Figure 1a

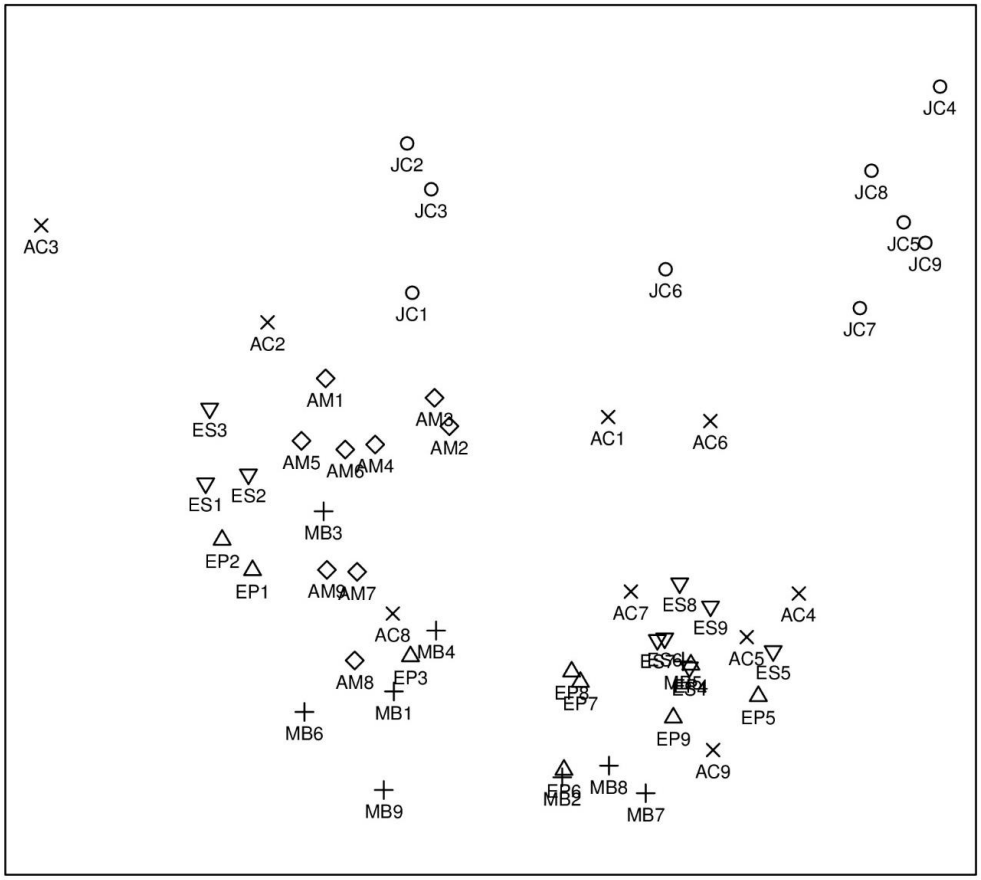

Figure 1b

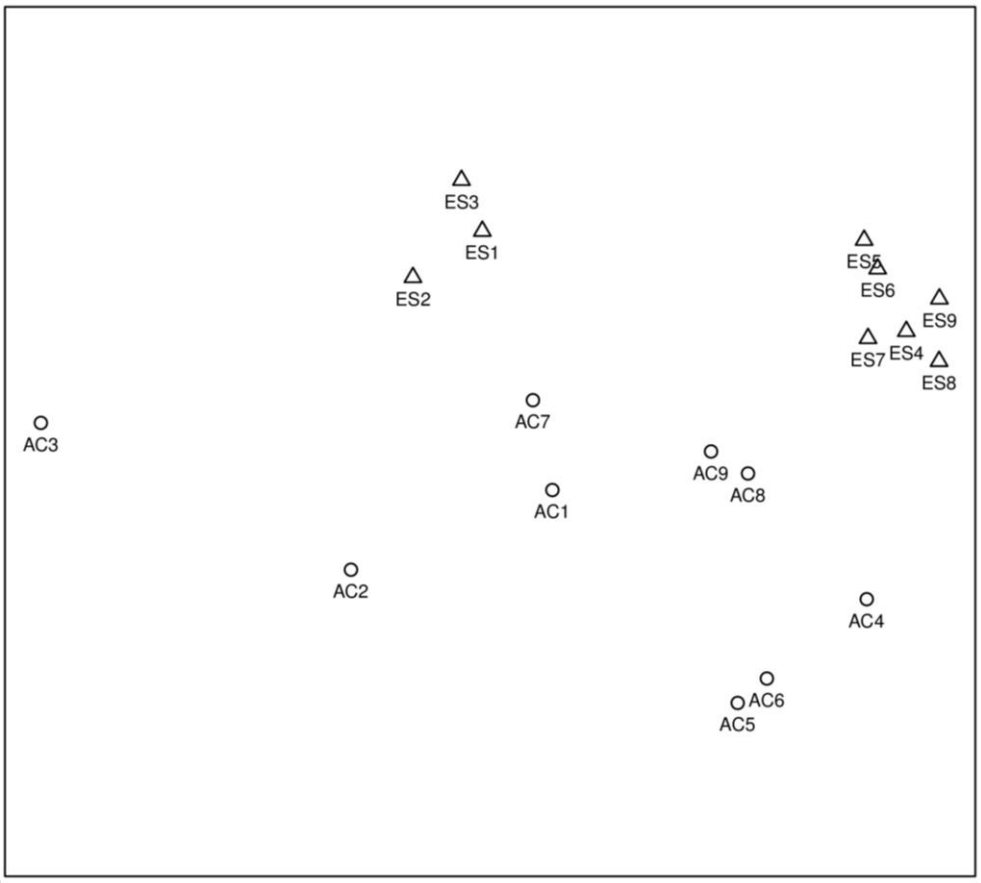

Figure 2a

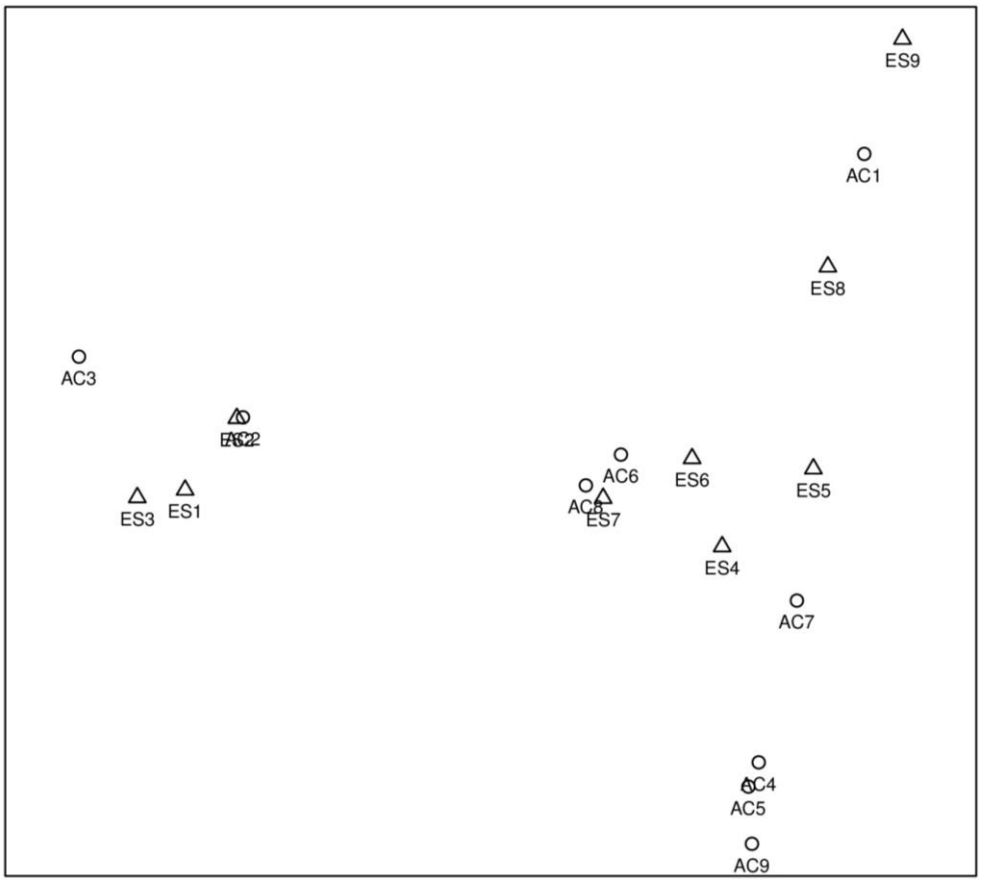

Figure 2b

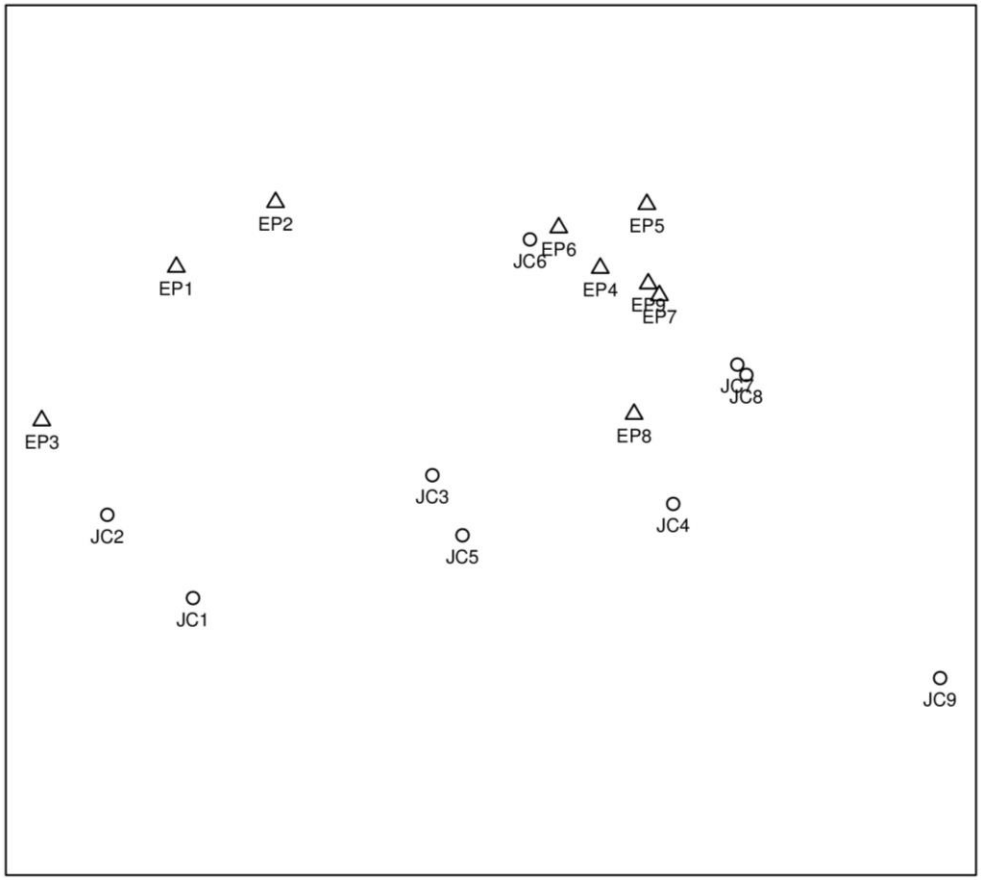

Figure 3a

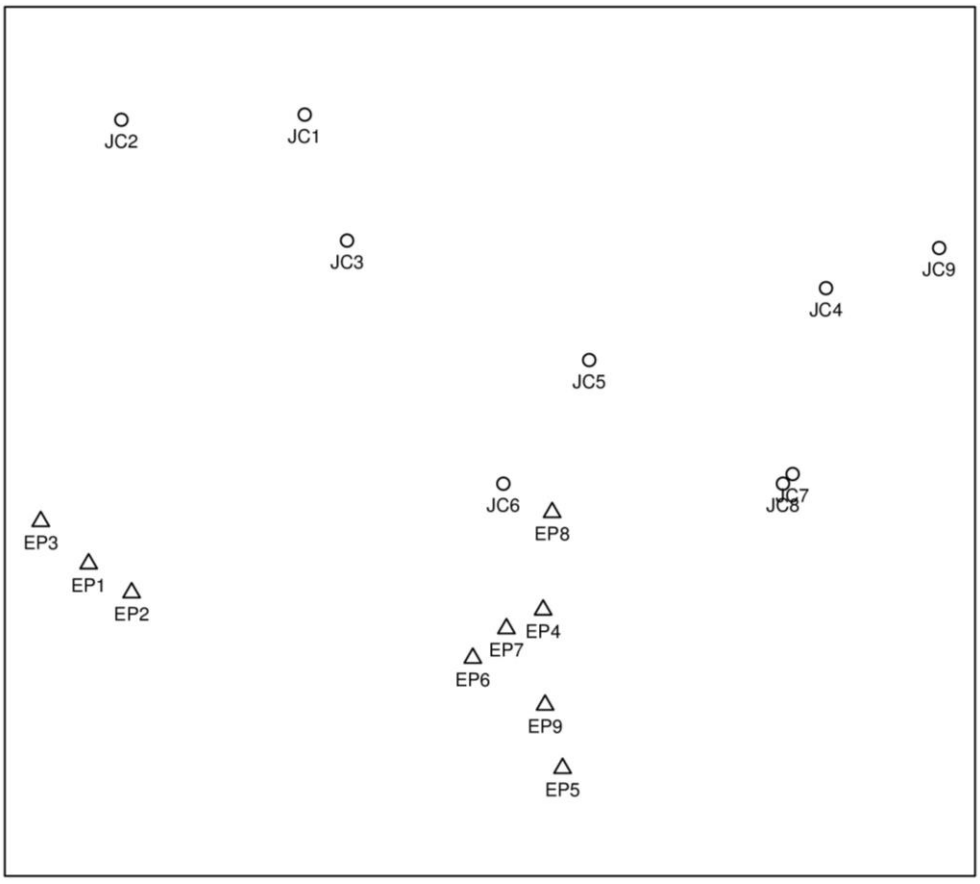

Figure 3b

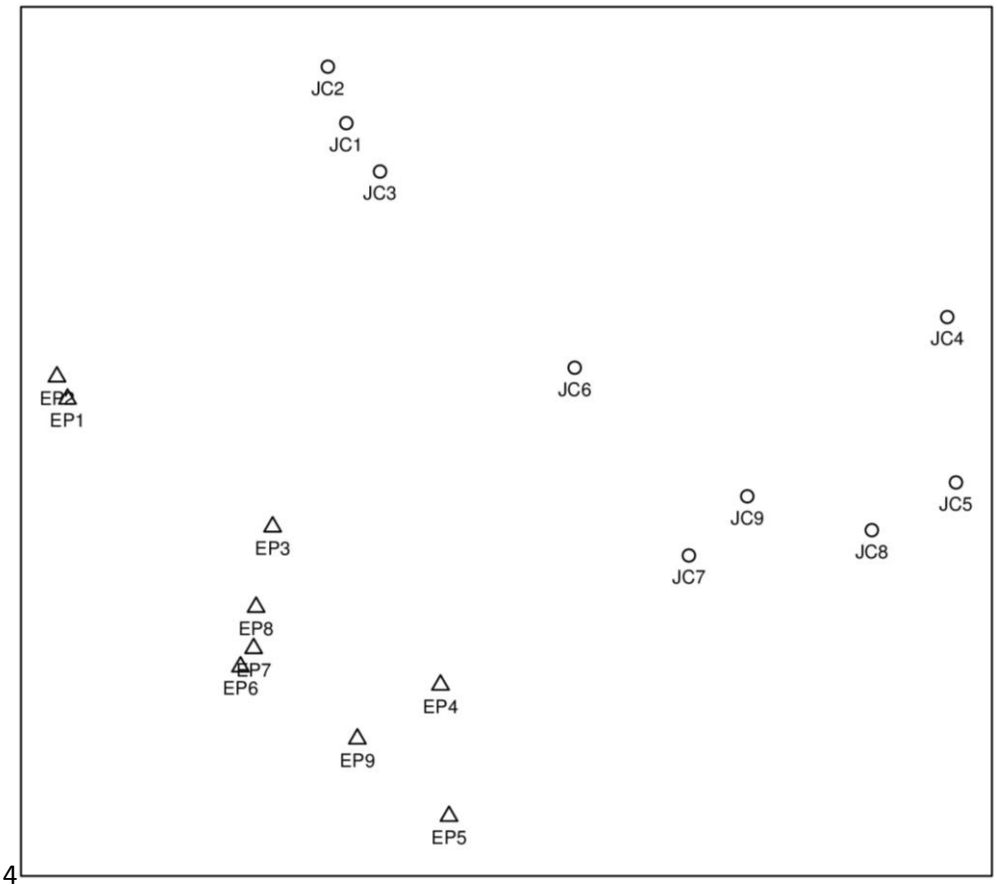
JC2
JC1
JC3
JC4
JC6
EP2
EP1
JC5
JC9
EP3
JC8
JC7
EP8
EP7
EP6
EP4
EP9
EP5

Figure 4

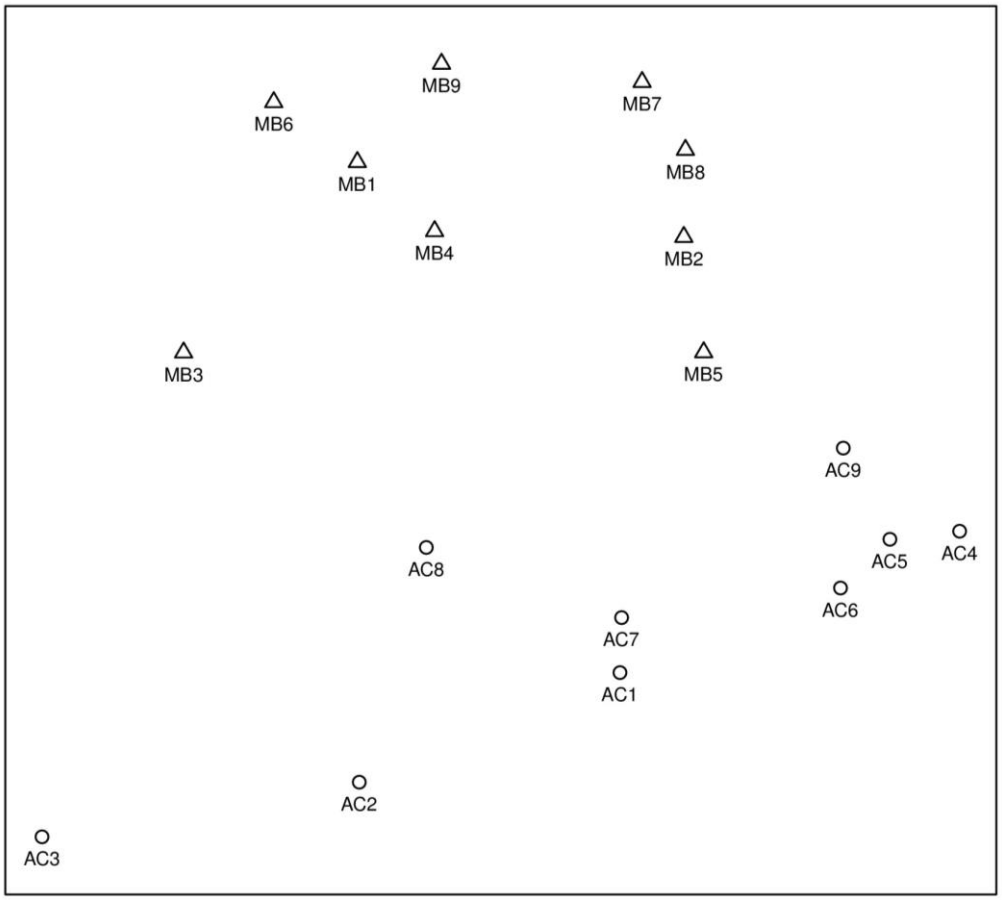

Figure 5a

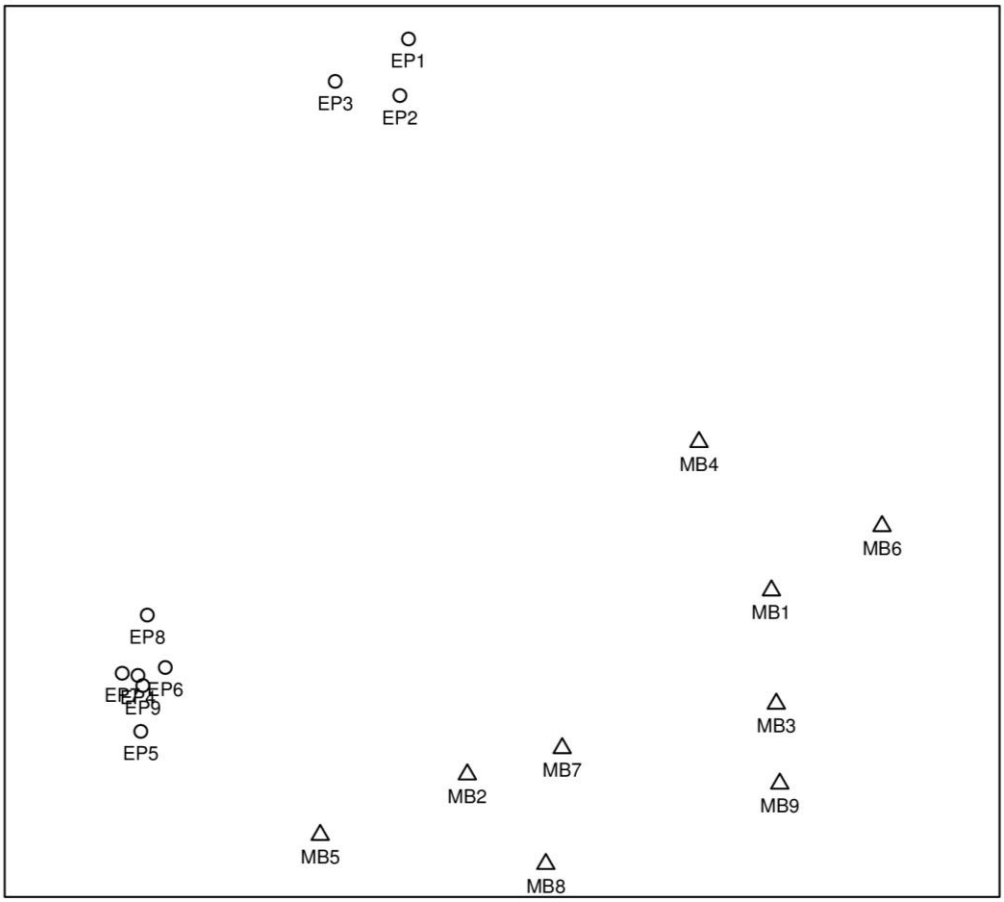

Figure 5b

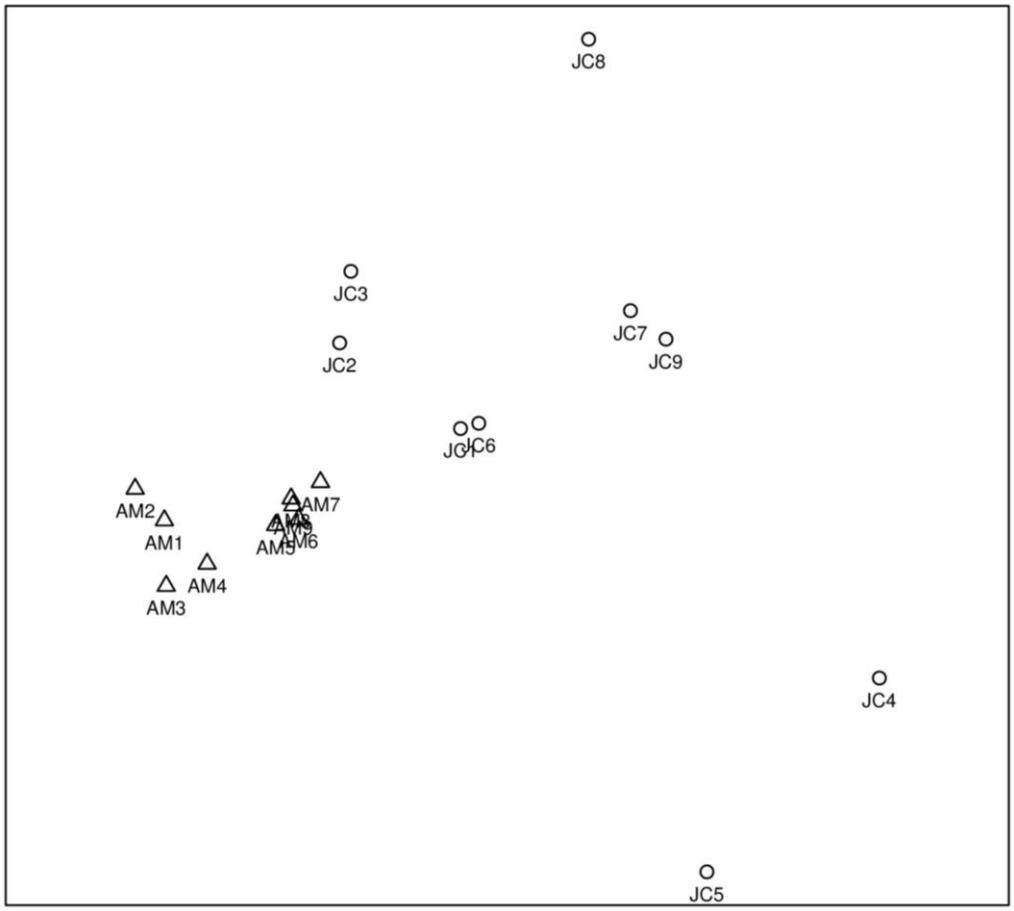

Figure 6a

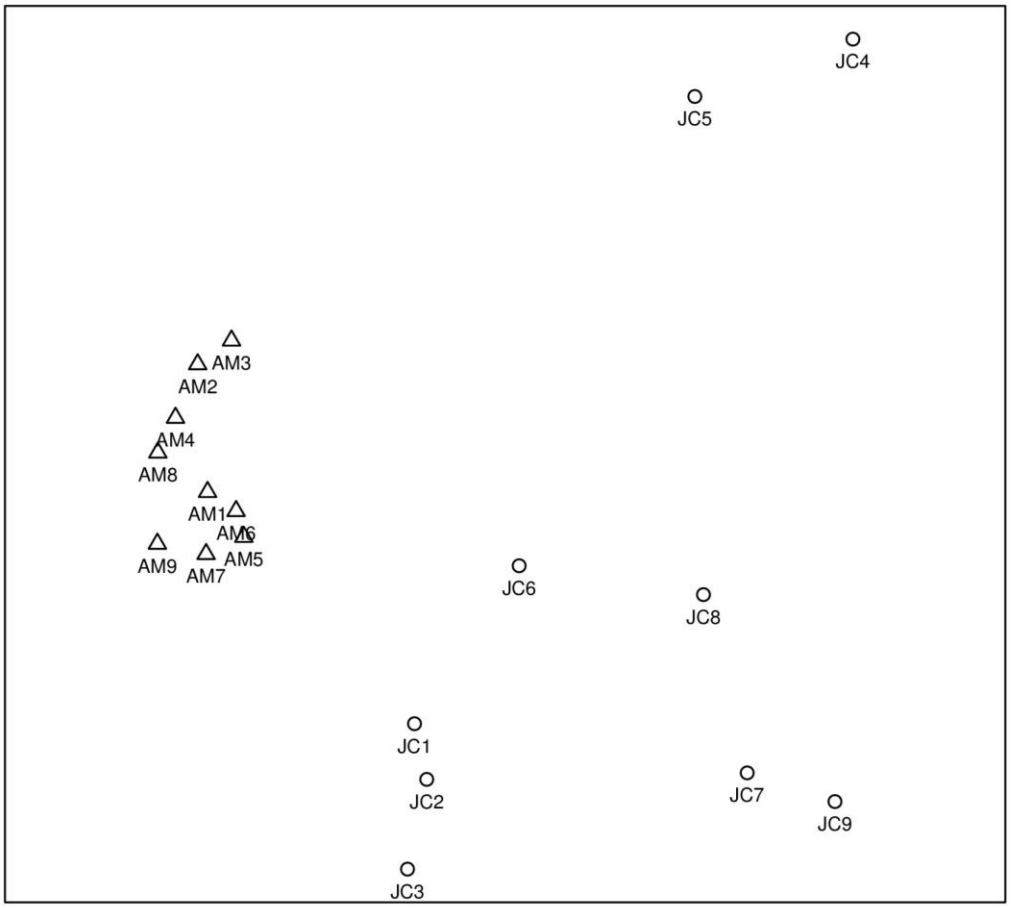

Figure 6b

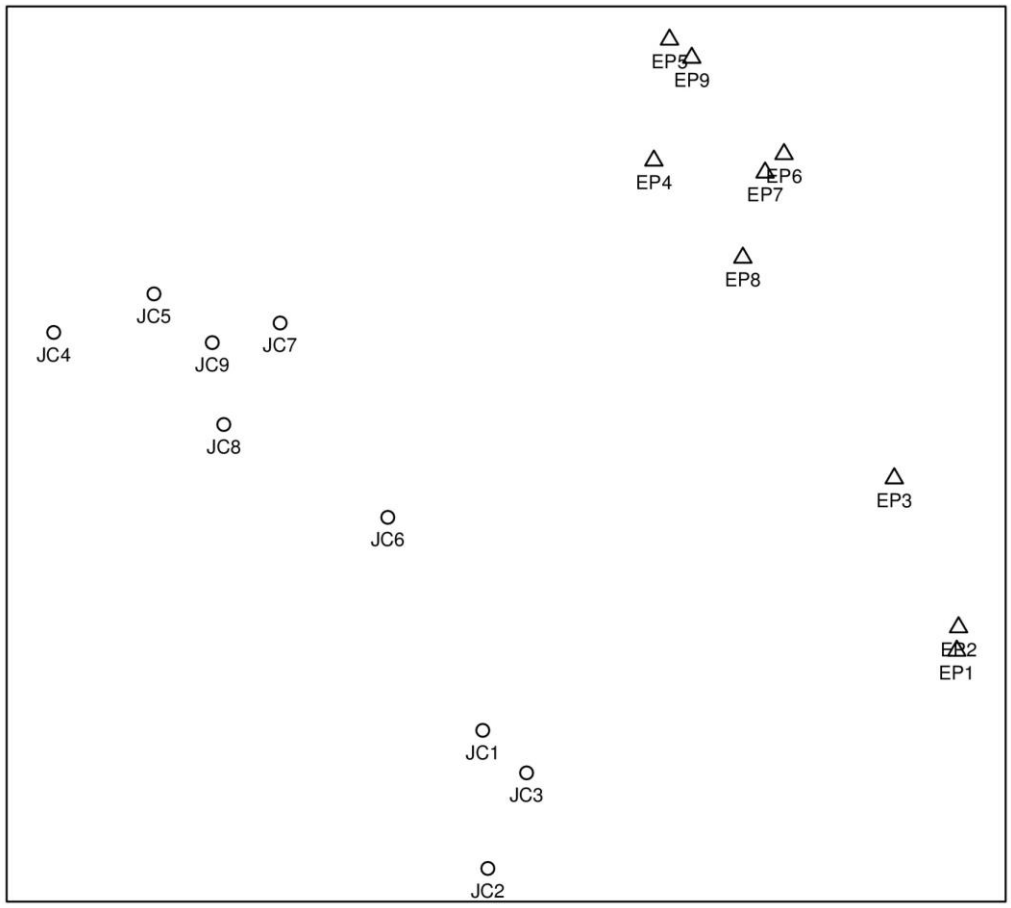

Figure 7

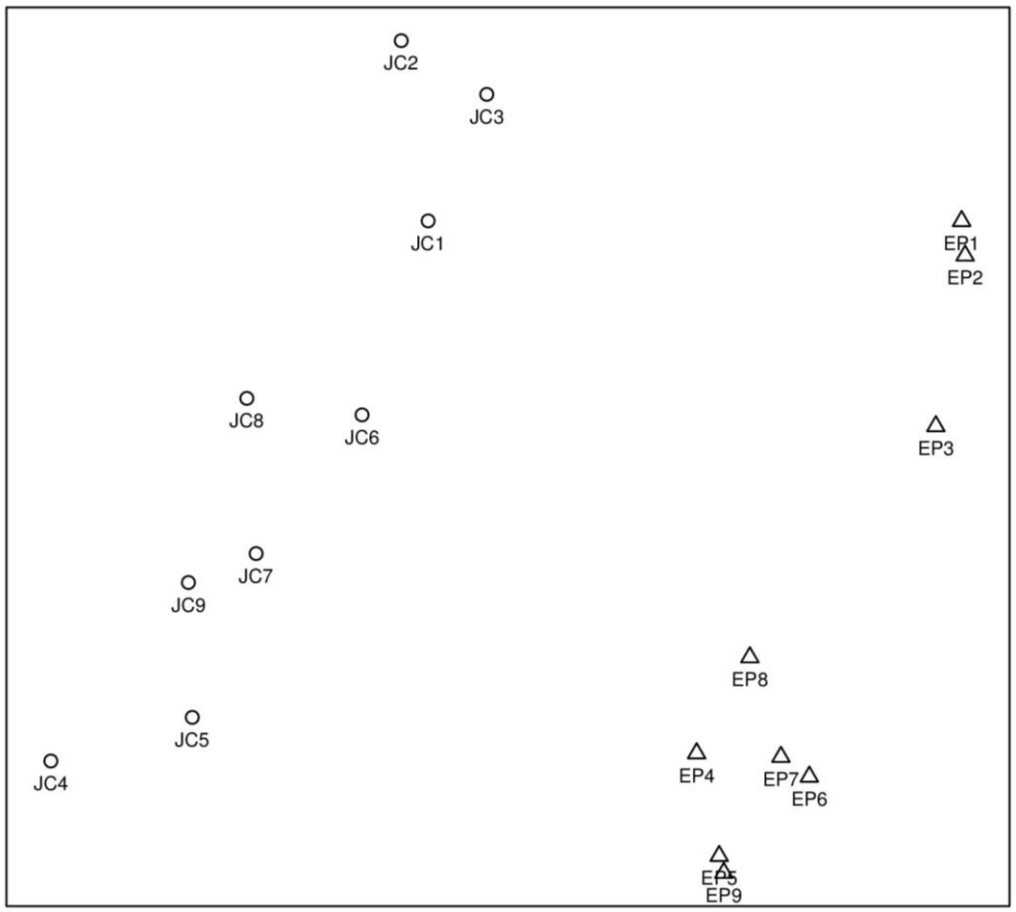

Figure 8