

## Smart Agro-Services in the Farm-Oriented Open Data in Europe (FOODIE) Platform: Insights from the Automatic Delimitation of Management Zones

Rodolfo B. Arango<sup>1</sup>, Cristina Monteserín<sup>2</sup>, Ismael S. Cerezo<sup>3</sup>, Antonio M. Campos<sup>4</sup>

### INFO

Received

Accepted

Available on-line

Responsible Editor:

### Keywords:

precision agriculture, smart services, management zones, decision support, FOODIE.

### ABSTRACT

Farm-Oriented Open Data in Europe (FOODIE) will provide specific and high-value applications and services for the support in the planning and decision-making processes of farmers and other stakeholders groups related to the agricultural and environmental domains. This paper focuses on FOODIE services based on machine learning algorithms and other artificial intelligence techniques providing, for instance, forecasting models to predict crop diseases or proposing management zones by means of clustering homogeneous zones of land. In particular, the service for the automatic delimitation of management zones is reviewed and tested in a case study. The results show the great potential of this smart agro-service for land classification.

## 1. Introduction

Smart services require huge amounts of data in order to train the models (Vapnik 2000) transforming data into valuable information. In particular, the agriculture domain deals with satellite imagery, agro-meteorological sensor networks, weather forecast, measurement of current soil and crop parameters or even data provided by farmers. However, such volume of data are not always available or easily accessible. Hence, platforms are needed to gather, unify and analyze these data converting them into valuable information, not only for experts but also for farmers. To that end, the FOODIE project (Farm-oriented Open Data in Europe 2014) is building an open and interoperable agricultural specialized platform hub on the cloud (Figure 1). This platform is intended to provide specific and high-value applications and services for supporting the planning and decision-making processes of different stakeholders groups related to agricultural and environmental domains.

Among these services, automatic land delimitation is one of the most challenging ones. In fact, the identification of homogeneous zones of crop land areas is a key factor (Schepers et al. 2004). These management zones (MZs) (Ferguson, Lark & Slater 2003) address spatial variability of crops grouping areas that share similar soil properties in order to apply specific farming practices to each MZ. For instance, farmers could apply site-specific treatments providing variable-rate fertilization (Yu et al. 2010) based on the current needs of the soil, the conditions of the crops and its phenological stage.

Although the knowledge of the farmer about crops and soil could be a starting point for MZs identification, other systematic approaches are required. In particular those more time- and cost-efficient methodologies that does not require exploring the whole parcel with any special equipment or taking samples of the soil for laboratory analysis. To this aim, we present the techniques considered by the FOODIE smart agro-service for the automatic delimitation of MZs. The remainder of the paper is organized as follows. Section 2 provides an overview of the FOODIE service platform. Section 3 describes the FOODIE smart agro-service for the automatic delimitation of MZs. Finally, Section 4 shows the conclusions and future work.

<sup>1</sup> Consultancy and Software, Seresco, Asturias, Spain  
Rodolfo.debenito@seresco.es

<sup>2</sup> Consultancy and Software, Seresco, Asturias, Spain  
Cristina.monteserin@seresco.es

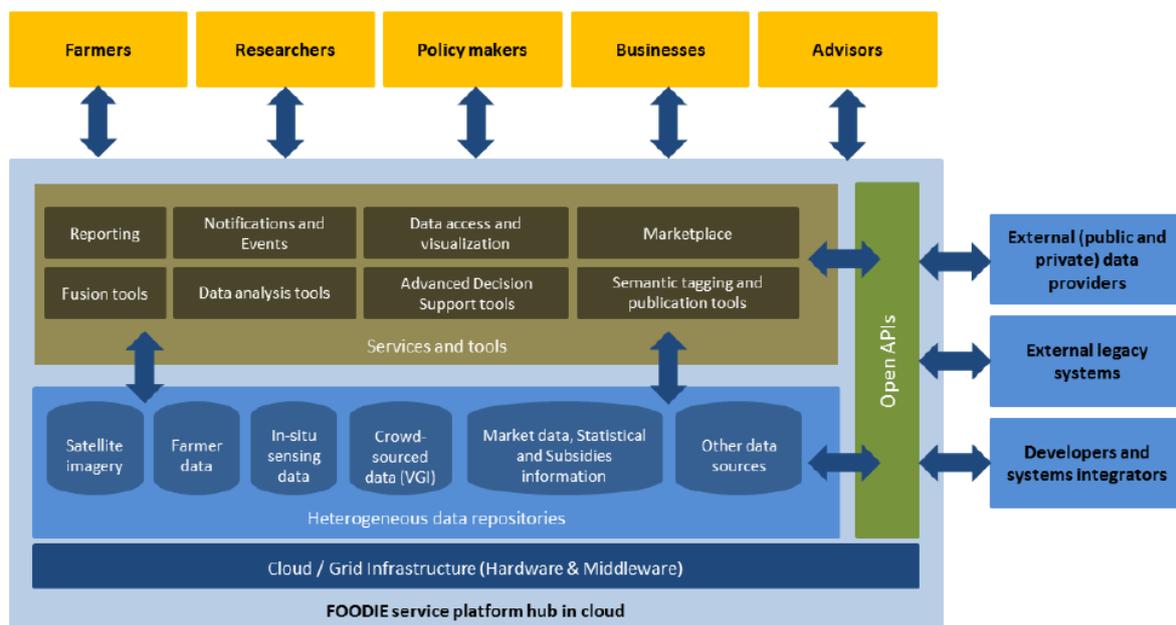
<sup>3</sup> Technology and Innovation, Seresco, Asturias, Spain  
Ismael.suarez@seresco.es

<sup>4</sup> Computer Science Department, University of Oviedo, Spain  
campos@uniovi.es

## 2. FOODIE. Open and Interoperable Agricultural Specialized Platform Hub

FOODIE project aims at providing a service platform in the cloud, which is built on principles of access and use of open data, interoperability and use of standards that supports the decision-making activities of different stake-holders groups across Europe in the agricultural area. Thus, the cloud based approach will seriously decrease the investment necessary for accessing to spatial data and services for all these different communities and it will help to business and governmental organisation in utilisation of spatial data. The project FOODIE aim at accomplishing the following technological objectives:

- To make use of existing spatial information resources and services for various Copernicus domains coming from different initiatives like INSPIRE (Directive, 2007/2/EC) or GMES/Copernicus (Copernicus 2015).
- To design and provide an open and interoperable geospatial platform hub on the cloud based on existing software components from research results and available solutions in the market.
- The integration of external agriculture production and food market data using principles of Open Linked Data.
- The inclusion of open and flexible lightweight Application Programming Interfaces (APIs), that will allow private and public stakeholders in the agricultural and environmental area to publish their own datasets and make them available in the platform hub as open linked data.
- To provide specific and high-value applications and services for the support in the planning and decision-making processes of the different stakeholders groups.
- The provision of security mechanisms to prevent the unauthorised access and use of the platform users' personal information as well as the data published by them.
- To provide a marketplace where data can be discovered and exchanged but also external companies can publish their own agricultural applications based on the data, services and applications provided by FOODIE.



**Figure 1.** FOODIE service platform hub in cloud

### 3. FOODIE Smart Agro-Service for the Automatic Delimitation of MZ

Computer services for the agriculture domain could be improved by means of machine learning algorithms and other artificial intelligence (AI) techniques related with prediction, classification and clustering. FOODIE platform could provide smart services intended to recommend the use of phytosanitary treatments using forecasting models to predict mildew, oidium or botrytis diseases, whilst other ones will be focused on yield prediction or on providing recommendations related with needs of irrigation (see Table 1). Among these smart agro-services, the automatic clustering of land areas (Figure 2) sharing similar soil properties (Ferguson, Lark & Slater 2003) is a key factor (Schepers et al. 2004) in order to apply specific farming practices. Although the knowledge of the farmer about crops and soil could be a starting point for MZs identification, other systematic approaches are required. In particular those more time- and cost-efficient methodologies that not require exploring the whole parcel with a specific equipment or taking samples of the soil for laboratory analysis.

**Table 1.** Examples of smart agro-services considered by the FOODIE platform

	FOODIE Main Source
Automatic clustering of crop land areas	Soil variability predicted by means of remote sensing
Site-specific mildew disease treatments	Forecasting of risk of mildew disease
Site-specific oidium disease treatments	Forecasting of risk of oidium disease
Site-specific botrytis disease treatments	Forecasting of risk of botrytis disease
Site-specific fertilization needs	Vegetation indices and remote sensed data
Site-specific irrigation needs	Moisture indices and vegetation indices
Recommendations about harvesting	Maturity indices and weather forecast
Yield forecasting	Historical records of weather conditions, past yields, soil nutrients and weather forecasts

Automatic zoning has been largely studied and researches had proposed various methods such as the classification of apparent soil electrical conductivity (Johnson et al. 2003) (Peralta & Costa 2013) or the analysis of yield maps (Blackmore 2003). Referring to previous researches and the type of features considered in the automatic delimitation of MZs, the methods could be classified as follows:

- **Soil properties.** In order to identify MZ, chemical properties of the soil were compared (Ortega & Santibáñez 2007) with several techniques such as PCA and cluster analysis. Supervised classification algorithms were tested with different datasets including soil maps, digital elevation models and apparent soil electrical conductivity (Simbahan & Dobermann 2006). In the same line other authors considered soil properties (Moral et al. 2011) (Fu, Wang & Jiang 2010) and also yield and crop quality (Aggelopoulou et al. 2013).

- **Topographical characteristics.** Homogeneous zones of a cotton field were identified (Schuster et al. 2011) considering two datasets: the first one with two estimators of yield and the second one considering geo-referenced field properties such as topographical characteristics and treatments applied to the field.

- **Biophysical features.** The use of biophysical features such as annual moisture deficit/surplus and mean annual precipitation was proposed (Liu & Samal 2002). The authors tested the same dataset with k-means and fuzzy algorithms concluding that fuzzy approach generates more accurate delineations.

- **Remote-sensed data.** Other approaches (Kumar et al. 2011) studied the use of the k-means algorithm with the MODIS-based greenness index and the seasonal leaf area index, developing a parallel implementation been able to delimit 1,000 agroecozones in 700 seconds using 2,048 processors.

FOODIE smart service for the automatic identification of MZs relies in a clustering algorithm and remote-sensed data from Landsat 8 satellite. The following subsections explain the techniques used by the service.

### 3.1. Data Input

Landsat 8 (Landsat 2015) satellite provide radiometric data valuable for precision agriculture (Villajos et al. 2008). This satellite takes about 400 images every day using the instruments: Operational Land Imager (OLI) and Thermal Infrared Sensor (TIRS). The first one collects data from 8 spectral bands with a spatial resolution of 30 meters and from a panchromatic band of 15 m. The second one offers two thermal bands of 100 m. resampled to 30 m. All the data collected by these instruments are publicly available in GeoTIFF format via web portals such as Earthnet Online (Earth Online 2015) or EarthExplorer (EarthExplorer 2014).

In order to generate the data input for the MZs identification process, we consider the raw values of the thermal infrared band B10 from TIRS data product. In particular, the values of that band for the extension of the case-of-study plot during a period of time of four months. These value are normalized and provided as input to the cluster algorithm. Next subsection details the clustering algorithm.

### 3.2. Clustering Algorithm

As it was aforementioned, the data input for the algorithm correspond to remote-sensed thermal infrared values of the considered plot during a four-month period. To obtain the automatic delimitation of the MZs, we used a partition clustering algorithm called Partitioning Around Medoids (PAM) using Manhattan distance as metric of similarity. The main characteristics of PAM (Li 2009) are the following:

- It is a partitioning algorithm. Thus, it breaks the input data up into groups until some stability condition is reached.
- The number of groups is defined in advance.
- PAM stands for “Partition Around Medoids”. It tries to find a set of objects called medoids that are centrally located in clusters.
- PAM is an algorithm more robust than K-means because it minimizes a sum of dissimilarities instead of a sum of squared euclidean distances.

The main difference between this algorithm and the classical K-means method is that PAM uses medoids as centers of the clusters and these medoids are selected among the objects to be clustered. The algorithm involves three steps:

1. Initialization. Select, at random, the  $k$  medoids from the data points
2. Assignment. For each point, locate the closest medoid and assign it to the corresponding cluster
3. Update. For each cluster, compute the new medoid from the points assigned to the cluster. The new medoid will be the point that minimizes the dissimilarity to the rest of the elements in the cluster.

Steps 2 and 3 are repeated until the clusters are no longer modified. Once the number of clusters is computed, the cluster assignment is retrieved, providing the land delimitation.

### 3.3. Selection of K clusters

A large number of ways of evaluating the goodness of a clustering algorithm have been proposed in the literature. In fact, the evaluation of clustering structures is the most difficult task in clustering algorithms. In this case, we selected the Silhouette coefficient (Rousseeuw 1987) and the Calinski-Harabasz index (Caliński & Harabasz 1974) methods in order to validate the clustering.

The Silhouette coefficient is based on the comparison of cluster tightness and separation. This Silhouette shows which objects lie well within their cluster, and which ones are merely somewhere in between clusters. The average silhouette width provides an evaluation of the clustering validity, and can be used to select an ‘appropriate’ number of clusters.

On the other hand, Calinski-Harabasz index evaluates the cluster validity considering the error sum of squares between different clusters and the squared differences of all objects in a cluster from their respective cluster center. The index is calculated as follows (Maulik & Bandyopadhyay 2002):

$$CH(K) = \frac{(\sum_{k=1}^K n_k \|z_k - z\|^2)/(K-1)}{(\sum_{k=1}^K \sum_{i=1}^{n_k} \|x_i - z_k\|^2)/(n-K)}$$

where  $n_k$  is the number of points in cluster  $k$  and  $z$  is the centroid of the entire data set.

Therefore, to select the optimum  $k$  according to the Silhouette coefficient and to the Calinski-Harabasz index, we use an implementation (Hennig 2010) of these methods and we consider both values of  $k$ . The implementation follows a procedure similar to the one described below in order to select the best value of  $k$  using the Silhouette coefficient (suppose that the number of points to cluster is  $n$  and that  $K^*$  is the maximum number of clusters, which is equal to or less than  $n$ ):

**for**  $j = 1, K^*$  **do**

**for**  $i = 1, n$  **do**

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}},$$

$$a(i) = \frac{\sum_{j \in C_i} (d(i, j))}{\#C_i}, \quad C_i \text{ cluster of element } i$$

$$b(i) = \min_C d(i, C), \quad \forall C$$

**end for**

$$s_{avg}^j = \frac{\sum_{i=1}^n s(i)}{n}$$

**end for**

$$k = \operatorname{argmax}\{s_{avg}^j\}$$

A similar procedure is followed by Calinski-Harabasz implementation, computing the index for each  $k$  and selecting the  $k$  with the highest index.

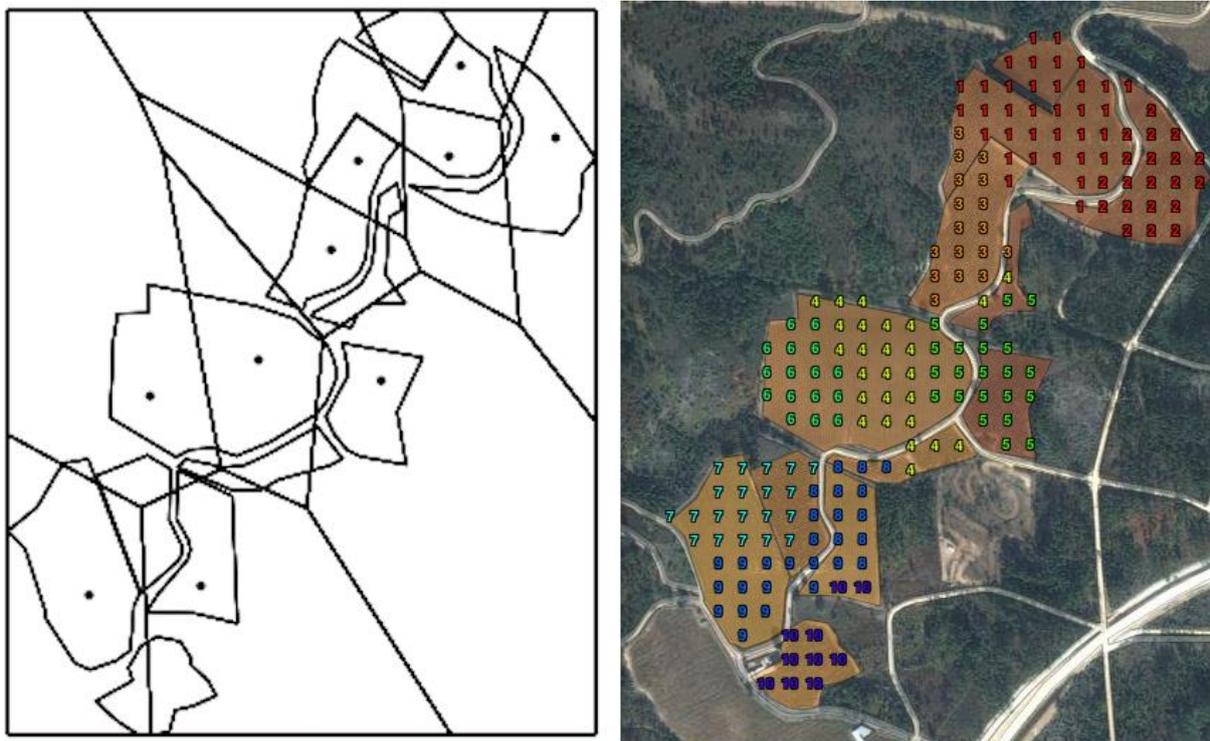
### 3.4. Evaluation and Experimental results

In order to verify the validity of the smart agro-service, we applied it to a case of study for the vineyards of Terras Gauda, a well-known wine producer in Galicia, Spain. Three different vineyard parcels with dissimilarities in topographical characteristics were considered. The outcomes were evaluated according to a previous study at this location for the delimitation of the MZs based on the experience and knowledge of the viticulture expert of Terras Gauda. The study considered factors such as yield production, topographical characteristics and soil properties in order to identify homogeneous zones of the vineyard and deploy an agro-meteorological station in each representative zone. However the study did not provide the contour of the MZs but the geolocation of the agro-meteorological stations. To address this lack of MZs delimitation, we calculated the extension of each MZ using a Voronoi partition (Voronoi 1908) considering the geolocations of the sensors (Figure 2). Then we compared the tessellation with the proposed clusters (Figure 2) by means of a confusion matrix (Kohavi & Provost 1998) in order to calculate the following metrics:

- True Positive Rate (TPR) or Recall. The percentage of positive cases that were correctly identified. Is calculated as  $\frac{d}{c+d}$
- True Negative Rate (TNR) or Specificity. The percentage of negatives cases classified correctly. Is calculated as  $\frac{a}{b+a}$
- Positive Predictive Value (PPV) or Precision. The percentage of the predicted positive cases that were correct. Is calculated as  $\frac{d}{b+d}$
- Accuracy (ACC). The percentage of spatial data points correctly classified. Is calculated as  $\frac{a+d}{a+b+c+d}$

Considering true negatives ( $a$ ) as the points correctly classified as land without vineyard, False positives ( $b$ ) as the number of points incorrectly classified as vineyard (even vineyard belonging to other nearby parcels outside Terras), False negatives ( $c$ ) as the points incorrectly classified as land without vineyard and True positives ( $d$ ) as the number of spatial data points correctly classified as vineyard.

The overall measurement results show an average of approximately 95% accuracy, 97% specificity, 73% recall and a 80% of precision. The results suggest that the use of PAM clustering algorithm with thermal infrared data from TIRS instrument is a promising method for the automatic MZs delimitation.



**Figure 2.** The figure at the left shows the Voronoi tessellation based on the actual location of the agro-meteorological stations, represented by dots, in a vineyard parcel. The figure at the right shows the proposed delimitation for the MZ based on the clustering algorithm. Each spatial data point is labeled with its corresponding cluster number.

#### 4. Conclusions

The results obtained in the case of study show the great potential of the FOODIE smart agro-service for MZ delimitation by means of PAM clustering algorithms with data collected from TIRS sensors of the Landsat 8 satellite. In particular, clustering the spatial data points of the land according

the values of the thermal infrared band B10. However, further research on how to optimize  $k$  for identifying the number of MZs is still required.

Future work will involve the optimization of  $k$  and the evaluation of the smart agro-service with other types of crops and countries.

## References

- Vapnik, V 2000. The nature of statistical learning theory. Springer.
- Farm-oriented Open Data in Europe, 2014. Available from: <<http://www.foodie-project.eu>>. [24 March 2015].
- Schepers, AR, Shanahan, JF, Liebig, MA, Schepers, JS, Johnson, SH, & Luchiari, A 2004 'Appropriateness of management zones for characterizing spatial variability of soil properties and irrigated corn yields across years', *Agronomy Journal*, vol. 96, no. 1, pp. 195–203.
- Ferguson, R, Lark, R, & Slater, G 2003 'Approaches to management zone definition for use of nitrification inhibitors', *Soil Science Society of America Journal*, vol. 67, no. 3, pp. 937–947.
- Yu, H, Liu, D, Chen, G, Wan, B, Wang, S, & Yang, B 2010 'A neural network ensemble method for precision fertilization modeling', *Mathematical and Computer Modelling*, vol. 51, no. 11, pp. 1375–1382.
- Directive, I. N. S. P. I. R. E. 2007. Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). *Official Journal*. 25 April.
- Copernicus 2015. Available from: <<http://www.copernicus.eu/>>. [28 May 2015]
- Johnson, CK, Mortensen, DA, Wienhold, BJ, Shanahan, JF & Doran, JW 2003, 'Site-specific management zones based on soil electrical conductivity in a semiarid cropping system', *Agronomy Journal*, vol. 95, no. 2, pp. 303–315.
- Peralta, NR & Costa, JL 2013, 'Delineation of management zones with soil apparent electrical conductivity to improve nutrient management', *Computers and Electronics in Agriculture*, vol. 99, pp. 218–226.
- Blackmore, S, Godwin, RJ & Fountas, S 2003, 'The analysis of spatial and temporal trends in yield map data over six years', *Biosystems engineering*, vol. 84, no. 4, pp. 455–466.
- Ortega, RA & Santibáñez, OA 2007, 'Determination of management zones in corn (*Zea mays* L.) based on soil fertility', *Computers and Electronics in Agriculture*, vol. 58, no. 1, pp. 49–59.
- Simbahan, GC & Dobermann, A 2006, 'An algorithm for spatially constrained classification of categorical and continuous soil properties', *Geoderma*, vol. 136, no. 3, pp. 504–523.
- Moral, F, Terrón, J & Rebollo, F 2011, 'Site-specific management zones based on the rasch model and geostatistical techniques', *Computers and Electronics in Agriculture*, vol. 75, no. 2, pp. 223–230.
- Fu, Q, Wang, Z & Jiang, Q 2010, 'Delineating soil nutrient management zones based on fuzzy clustering optimized by fPSOg', *Mathematical and Computer Modelling*, vol. 51, no. 11–12, pp. 1299 – 1305.
- Aggelopoulou, K, Castrignanò, A, Gemtos, T & Benedetto, DD 2013, 'Delineation of management zones in an apple orchard in Greece using a multivariate approach', *Computers and Electronics in Agriculture*, vol. 90, no. 0, pp. 119 – 130.
- Schuster, E, Kumar, S, Sarma, SE, Willers, J & Milliken, G 2011, 'Infrastructure for data-driven agriculture: identifying management zones for cotton using statistical modelling and machine learning techniques', in *Emerging Technologies for a Smarter World (CEWIT), 2011 8th International Conference Expo*, Long Island, New York, pp. 1–6.
- Liu, M & Samal, A 2002, 'A fuzzy clustering approach to delineate agroecozones', *Ecological modelling*, vol. 149, no. 3, pp. 215–228.
- Kumar, J, Mills, RT, Hoffman, FM & Hargrove, WW 2011, 'Parallel k-means clustering for quantitative ecoregion delineation using large data sets', *Procedia Computer Science*, vol. 4, pp. 1602–1611.
- Landsat 2015. Available from: <<http://landsat.usgs.gov/>>. [24 March 2015]
- Villajos, SO, Villar, AA, Peña, MM, Arroyo, MP, Alcázar, GV, Morera, JP & García, LP 2008, 'Los satélites de media y baja resolución espacial como fuente de datos para la obtención de indicadores ambientales' *IX Congreso Nacional de Medio Ambiente*, Madrid.

- Earth Online 2015, *Data Access*. Available from: <<https://earth.esa.int/web/guest/data-access>>. [24 March 2015]
- Earthexplorer 2014. Available from: <<http://earthexplorer.usgs.gov/>>. [24 March 2015]
- Li, X 2009, 'K-Means and K-Medoids' in *Encyclopedia of Database Systems*, eds Springer US, pp. 1588–1589.
- Rousseeuw, PJ 1987, 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, vol. 20, no. 0, pp. 53 – 65.
- Caliński, T & Harabasz, J 1974, 'A dendrite method for cluster analysis', *Communications in Statistics-theory and Methods*, vol. 3, no. 1, pp. 1–27.
- Maulik U & Bandyopadhyay, S 2002, 'Performance evaluation of some clustering algorithms and validity indices', *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, no. 12, pp. 1650–1654.
- Hennig, C 2010, 'fpc: Flexible procedures for clustering', *R package version*, vol. 2, pp. 0–3.
- Voronoi, G 1908, 'Nouvelles applications des paramètres continus à la théorie des formes quadratiques. deuxième mémoire. recherches sur les paralléloèdres primitifs', *Journal für die reine und angewandte Mathematik*, vol. 134, pp. 198–287.
- Kohavi, R & Provost, F 1998, 'Glossary of terms', *Machine Learning*, vol. 30, no. 2-3, pp. 271–274.