Exploring Urban Tree Site Planting Selection in Mexico City through Association Rules

Héctor Javier Vázquez[1] and Mihaela Juganaru-Mathieu[2]
[1]*Departamento de Sistemas, Universidad Autonoma Metropolitana, Unidad Azcapotzalco, Avenida San Pablo 180, Mexico D.F., Mexico*
[2]*Institut H. Fayol, Ecole Nationale Supérieure des Mines, 158, cours Fauriel, 42023, Saint Etienne, France*

Keywords: Data Mining, Association Rules, Prediction, Rule Validation, Urban Trees, Planting Sites.

Abstract: In this paper we present an exploration of association rules determine planting sites considering urban tree's characteristics. In first step itemsets and rules are generated using the unsupervised algorithm Apriori. They are rapidly characterized in terms of tree planting sites. In a second step planting sites are fixed as target values to establish rules (a supervised version of the a priori algorithm). An original approach is also presented and validated for the prediction of the planting site of the species.

1 INTRODUCTION

Now days there is no doubt about urban trees benefits for the urban population in comparison with the situation more than 25 years ago. Besides the development of different practical and theoretical studies to preserve environment and to maintain a sustainable city (Badii et al., 2008; Watson, 2011), high air pollution levels and climate change are among the main detonators to promote tree planting campaigns. In Mexico City, protection of trees has required important efforts from academics and different social groups worried about urban sustainability. Although, authorities claim that in Mexico City, trees are for all urban inhabitants and that maintenance and planting trees campaigns are almost all around the city; a more careful look still reveals, high and uneven trees presence, as they concentrate mainly in certain urban zones. In practice it is important to say that tree planting campaigns come most of the times after the economic interests of urban economic agents more interested, to build urban roads, commercial and residencies; as in many urban environments, historically, in the city of Mexico trees are mainly located in high-level economic areas and less present in low-income population neighbourhoods. This is not surprising, given the value that trees give to properties. High maintenance needs and low tree survival is mainly observed in low income areas even though different scientific studies shows that urban trees benefits are not just economical and visual, but essential for urban life. Luckily, in the last years, city's authorities have been more sensible to promote and to establish politics to plant, to protect and to increase trees survival of urban trees in most of the Mexico city's boroughs. However, these objectives are difficult to achieve considering, Mexico city's variety of conditions, places, climes and its huge population. For example, weather conditions are varied, some zones are cold others very dry and others rainy; trees can be found along or in sidewalks, in road medians, in gardens, parks, sport fields, and cemeteries. With this, it is common for urban trees to show low survival rates, to suffer severe injuries and diseases originated from their environment: air, soil and water pollution, insects, parasites, lack of water, impediments to growth, such as cabling, planting pits and soil compaction, and damage from vandalism. Diverse studies have been pursued to establish criteria to plant trees, among them, it has been considered trees species, tree's density and tree's diversity, nevertheless there exists consensus that site is one of the most important criteria to consider when planting trees in an urban environment (Kuhns and Rupp, 2000). It is known that good quality of sites increases probability of maintaining healthy trees and survival whereas low quality sites diminish tree's health and survival. In fact soil characteristics (such as soil contaminants, earth compaction, drainage and aeration), site characteristics and location (for example site size, air pollution, streets, sidewalks,

Vázquez, H. and Juganaru-Mathieu, M.
Exploring Urban Tree Site Planting Selection in Mexico City through Association Rules.
DOI: 10.5220/0006085604250430
In *Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016) - Volume 1: KDIR*, pages 425-430
ISBN: 978-989-758-203-5

parks, utility lines) and environment (like wind speed, temperature, population stress) seem to be factors linked to tree survival. Added to this, not all trees adapt to any site and some are more adequate for a given site than others. Therefore tree's planting guidelines should include site selection criteria, good practices for site preparation and tree selection (Longman et al., 1993). More specifically, however given the large number and variety of different site variables the task to choose a site for a given tree species is far from being easy. In the field, there is a strong need of "ready out of the box rules" to select a planting site for one or different species. In terms of the species characteristics, tolerance to environmental stress and pollution examples of rules would be: *If the specie is tolerant to dry weather, vandalism and to high pollution levels then the tree can be planted on a side walk.*

In this work, association rules (Agrawal et al., 1993), are proposed to discover relevant characteristics and their relation with the planting site. Section 2 presents our data. Section 3 presents a study of items (frequent, maximal and closed), rules generated for a given site and results are compared with a Multiple Correspondance Analysis. Finally, conclusions and future work plans are presented in section 6.

2 DATA SET

Data was obtained from two technical manuals from the environmental secretary of Mexico City proposing guidelines for pruning, felling and transplanting trees and shrubs (SMA, 2000) and information about environment and pollution in Mexico city (SMA, 2000).

2.1 Data Set Description, Cleaning, and Preparation

This data was already presented in a previous article (Vazquez and Juganaru-Mathieu, 2014). This data can be obtained from the following address http://www.emse.fr/∼mathieu/data/trees/. For the present article, it seems important to recall that variables are organized in five groups:

- *Group 1*: *species name*, *genus*, *origin*
- *Group 2*: *tree*, *shrub*, *palm*, *fruit*, *evergreen*
- *Group 3*: tolerance to environment: to cold ($tcold$), to dryness ($tdry$), to mistreatment ($tmiss$) and to soil salinity ($tsal$)
- *Group 4*: recommended planting sites: streets and middle-roads (s_street), urban recreational parks (s_urbrp), parking lots (s_parlot), beneath electric lines ($s\_beleclin$), cemeteries (s_cem), sport fields (s_sportf), urban forest (s_urbfor)
- *Group 5*: sensitivity to air pollution levels: $veryhighpollution$, $highpollution$, $mildpollution$ and $lowpollution$.

The 134 species includes 65 different genera: 72.39% are trees, 50% are shrubs, 4.48% are palms, 11.94% are fruit tree, 66.42% are evergreen, 82.84% resist cold, 49.25% resist dryness, 32.09% tolerate soil salinity, and 29.10% are tolerant to mistreatment. Concerning planting sites almost all trees are recommended for urban parks (97%), for sport fields (89.5%), for cemeteries (86.5%), for parking lots (78.3%), for streets and middle roads (52.9%) and to be planted below electric lines (50.7%). These characteristics are not exclusive, as the same species may share more than one characteristics and the same tree might be proposed to be planted in more than one site. In the case of species response to pollution, a given species might have the same response to one or more pollution levels (if a given species is resistant to high pollution, it can be also resistant or adapted to the next lower level of pollution.

2.2 Data Cleaning and Preparation

We presented missing data estimation in a previous study (Vazquez and Juganaru-Mathieu, 2014), so in this work data without missing values are used.

The initial data are considered as a transactions record; each row corresponds to a tree (a transaction). An item is the value set to an attribute, for example, lowpollution=yes or evergreen=no. In the case of attribute genera, the value will be one of the 65 different genera considered. Each transaction is described by a set of items, so the database includes 134 species, each with 21 attributes.

Next section will give more details about the database, items, transactions, item sets and association rules assessment.

3 EXPLORING ASSOCIATIONS

Exploring associations can be pursued through the study of itemsets and rules. To facilitate understanding of results, lets consider a brief review of some definitions (Zaki and Wagner Meira, 2014).

Let define $I = \{x_1, x_2, \ldots, x_m\}$ as a set of m elements called items. An itemset is subset X of I,

$X \subset I$. An itemset of cardinality (or size) k is called a k-itemset. The database $\mathcal{D} = \{t_1, t_2, \ldots, t_n\}$ is a set of transactions, identified by an identifier; each transaction contains a subset of the items described in I.

To evaluate an itemset, the support $sup(X, \mathcal{D})$ is the number of transactions in the database D that contain the given itemset X. If $sup(X, D) minsup$, the itemset X is frequent. If $X \subset Y$, Y is a superset of X; if Y is frequent, X is also frequent. A frequent itemset X is called maximal if it has no frequent supersets. A frequent set X is closed if it has no frequent superset with the same support.

From frequent itemsets the association rules are obtained comparing items' frequencies. A rule, constructed with several items, has the form

$$X = (x1, x2, , xj) \rightarrow Y = (y1, y2, , yk)$$

where, X and Y are itemsets in I and are disjoint $X \cup Y = \emptyset$; X and Y are called antecedent (left-hand-side or LHS) and, respectively, consequent (right-hand-side or RHS) of the rule (Hahsler et al., 2005). Each rule then evaluated using concepts such as support and confidence.

Support of a rule is defined with conjoint probabilities $P(X \cap Y)$, as the fraction of transactions that contain both X and Y and confidence of a rule is defined as the conditional probability $P(Y|X)$ which measures how often the items of Y appear in transactions that contain X.

$$confidence(X \rightarrow Y) = P(X|Y) = \frac{support(X \cup Y, \mathcal{D})}{support(X, \mathcal{D})}$$

Given a set of transactions $\mathcal{D}$, the goal of association rule mining is to find all rules having the support up to *minsup* (a threshold) and the confidence up to *minconf* (an other threshold). Positive correlation between Y and X of rule $X \rightarrow Y$, also called the lift is defined as

$$lift(X \rightarrow Y) = P(Y|X)/P(Y) = \frac{P(X \cap Y)}{P(X)P(Y)}$$

Further details about all these concepts and operations can be found in (Zaki and Wagner Meira, 2014).

To obtain the different itemsets (frequent, closed and maximal), to generate and to explore rules, input data can be just a $n \times m$ table with n species on lines and m attributes on columns or a list of transactions. In all cases, a sparse matrix is produced. R program (R Core Team, 2014) and R libraries `arules` (Hahsler et al., 2005) and `arulesViz` (Hahsler and Chelluboina, 2011).

4 RESULTS

4.1 Items and k-Itemset Focus

We realize a first exploration considering the attributes retained from Groups 1, 2, 3, 4 and 5. From the transactions records we obtain the transactions set as an itemMatrix in sparse format with 134 rows (transactions) and 266 columns (items): 134 species, the 65 genera, the 27 origins, and the 10 different levels for pollution and 30 attributes (yes and no values) for the other variables. As many entries are empty, density of the matrix is 0.0789, meaning that only 2814 entries out of 35644 contain a value.

The absolute frequency distribution of items related to recommended and not recommended (yes and no values) planting sites is: streets and middle-roads (71, 63), urban recreational parks (130, 4), parking lots (105, 29), beneath electric lines (73, 61), cemeteries (116, 18) and sport fields (120, 14).

With the Apriori algorithm k-itemsets can be generated *starting with 1-itemsets*. However it is necessary to fix a minimum support. If, for example, a minimum support of 0.01, is fixed, 10525099 itemsets are generated. This huge amount of item sets is not directly exploitable and requires to look for other strategies such as to increase the threshold for minimum support or to obtain subsets containing a given item of interest. Fixing a threshold of 0.1 for minimum support, we obtain 272958 frequent itemsets (with a minimum of 476 itemsets with 2 items and a maximum of 22647 itemsets with 10 items), 38294 closed frequent itemsets and 23771 maximally itemsets (with a minimum of 2 itemsets with 2 items and a maximum of 22647 itemsets with 10 items).

Considering, that our aim is to predict plantation sites (variables from Group 4), it is important to evaluate the distribution of itemsets containing planting sites and the more adequate species or genus (variables from Group 1). The distribution of itemsets containing a least one planting site is presented in Figure 1.

Table 1 presents, for a minimum support of 0.1 and 0.15, the distribution of the number of itemsets (frequent, closed and maximally) and the number of item sets that includes genus Pinus and Quercus.

This exploration of item sets can continue: increasing or diminishing minimum support or eliminating items with lower frequencies. If minimum support is reduced this may not be enough to reveal attributes from Group 1 (Table 2), given their low frequencies. Another option to consider is to reduce the number of items, eliminating those with

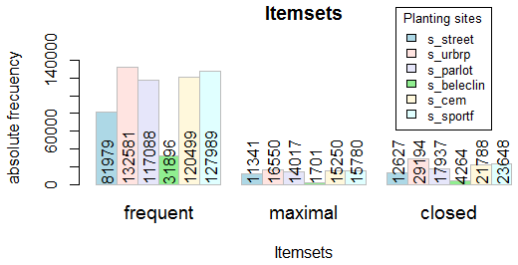

Figure 1: Distribution of itemsets (*minsup* = 0.1) containing at least 2 items related to the Planting Sites.

Table 1: Distribution, with minimum support of 0.1 and 0.15 for the number of itemsets (frequent, closed and maximal) for items Pinus, Quercus.

Itemsets	MinSupp	Frequency	Pinus	Quercus
Frequent	0.1	272958	10086	63
	0.15	82964	767	0
Closed	0.1	38294	511	1
	0.15	12321	3	0
Maximal	0.1	23771	510	1
	0.15	3359	2	0

lower frequencies, for example to eliminate items from Group 1 and just keeping items from Group 2 to 5. Items from Group 1 can be considered as supplementary, even though these items are not used to generate rules, they are linked to the item sets generated with Groups 2 to 5. In the next subsection will explore this option with 0.15 as minimum support.

4.2 Rules

Firstly, association rules are generated, with 0.15 as a minimum threshold for support, from the transactions records with just the 38 attributes values (yes, no and levels for pollution) from Groups 2, 3 and 5 in the antecedent and targeting in the right hand side both attribute values (yes and no) for each planting sites (Group 4). The number of rules generated are for each site are: (s_street, 8293), (s_urbrp, 35142), (s_parlot, 29799), (s_beleclin, 11717), (s_cem=yes, 30089), (s_sportf, 33523). Considering the high number of rules, different tests were realized changing restrictions on the lhs and rhs of the rules. Finally, for this exploration, affirmative attribute values are retained; as rules interpretation is, at first glance, easier. For each of the six planting sites the number of generated rules, the left hand side (lhs) of the rule retained ant its evaluations between parenthesis the triplet support, confidence and lift are:

- streets and middleroads: rules generated = 5; lhs of retained rule: tree=yes & tdry=yes & veryhighpollution=1 & mildpollution=3 (support = 0.20, confidence = 0.86, lift = 1.64).
- urban recreational parks: rules generated = 208; lhs of retained rule: veryhighpollution=1; (support = 0.24, confidence = 1.0, lift = 1.03).
- parking lots: rules generated = 134; lhs of retained rule: tdry=yes & veryhighpollution=1 & mildpollution=3; (support = 0.29, confidence = 1.0, lift = 1.27)
- beneath electric lines: rules generated = 52; lhs of retained rule: shrub=yes & tdry=yes & highpollution=3; (support = 0.20, confidence = 1.0, lift = 1.83)
- cemeteries: rules generated = 166; lhs of retained rule: veryhighpollution=1 & highpollution=3; (support = 0.19, confidence = 1.0, lift = 1.15)
- sport fields: rules generated = 195; lhs of retained rule: tree=yes & tcold=yes & tsalt=yes (support = 0.20, confidence = 1.0, lift = 1.11)

We observe an important reduction in the number of rules generated compared with the large number of item sets, mainly due to an increase of the threshold for minimum support, targeting just for affirmative items and for not considering items, from the group 1. Streets and middle roads is the site with less rules, whereas cemeteries, sport fields and urban recreational parks account for most of the rules. On the left hand side of the rules, the number of items varies between 1 for urban recreational parks to 4 for streets and middle roads. For the site beneath electric lines, shrubs are recommended, whereas for sport fields, trees have to be tolerant to cold and tolerant to salinity. All planting sites, but sport fields, fix as condition an item related to pollution. Species planted in streets and middle roads and parking lots should be at least resistant to mild pollution.

If all planting sites are considered, at the same time in the left hand side, 760 rules are generated Some findings are: all the rules for site s_beleclin show the highest lift, the number of items in the left hand side varies from 4 to twelve, item concerning sensibility to veryhighpollution is present in 145 rules and to resistance to middlepollution is in 45 rules. An inspection of all these 760 rules show that all retained rules. Most frequent genus recommended and the number of different species recommended for each planting site are:

- streets and middle-roads: most frequent genus: Pinus (13), Quercus (6), Cupressus (2). Number of different species = 27.
- urban recreational parks: most frequent genus: Pinus (22), Quercus (14), Prunus (4), Cupressus

(4), Ficus (3), Salix (3), Ulmus (3), Citrus(2), Alnus (2). Number of different species = 95.

- parking lots: most frequent genus: Pinus (14), Quercus (8), Cupressus (3), Juniperus (2). Number of different species = 40.
- beneath electric lines: most frequent genus: Pinus (5), Quercus (4), Juniperus(4), Ligustrum (3) Pittosporum (2). Number of different species = 27.
- cemeteries: most frequent genus: Pinus (8), Quercus (5), Ficus (3), Juniperus (3). Number of different species = 26.
- sport fields: most frequent genus: Pinus (8), Acacia (4), Eucalyptus (2). Number of different species = 27.

These results show that although Genus Pinus and Quercus are the most frequent, there is more genus variety recommended (number of different genus) in urban recreational parks, parking lots, cemeteries and sites located beneath electric lines. This last point is important considering that in Mexico city most of the electric and communication lines are aerial. These findings agree with the results obtained in previous works using other data mining methods such as Multiple Correspondance Analysis (Vazquez and Juganaru-Mathieu, 2014). However it is important to propose a method to validate these findings through a validation test. This will be explored in the next section.

5 ASSOCIATION RULES FOR PREDICTION

Generating association rules has firstly a descriptive aim and allow people to understand how to choose plating sites. On the other hand, our data collection contains only 134 species that clearly does not cover all the species in the area of Mexico City. Also it is high possible to try to plant some new trees, shrubs or palms from exotic origin. For all these considerations we are also interested to be able to predict which planting sites are adapted to a given specie, knowing basic characteristics, as the species presented in section 2. The main idea is to take only the rules having in the left side the characteristics of a new tree and having in right side only one attribute corresponding to a planting site. We will present a simple prediction algorithm based on a collection of association rules based on this simple idea and on some observations; we will validate this approach by a leave-one-out cross validation.

5.1 Algorithm

An association rule with the right side indicating planting site has the form :

$$Attr_1, Attr_2 \dots Attr_k \rightarrow site\_no \text{ or } site\_yes$$

a support s, a confidence $conf$, where site can be: s_street, s_urbrp, s_parlot, $s\_beleclin$, s_cem, s_sportf. We will name this form as "restricted planting" association rules. The support and confidence are significant if they are grater than some bounds min_supp and min_conf. Among other quality indicators (see (Lallich et al., 2007)) for an association rule we will take into account the lift.

If we have two association rules :

$$Attr_1, Attr_2 \dots Attr_k \rightarrow site\_yes$$
$$Attr'_1, Attr'_2 \dots Attr'_j \rightarrow site\_no$$

with $Attr_1, Attr_2 \dots Attr_k, Attr'_1, Attr'_2 \dots Attr'_j$ attributes characterizing a given new tree, we will "prefer" the rule with the higher lift. This means, that if the associations rules can infer an information and the negation of it, each one with a score, we will take into account the highest score.

If we have two association rules :

$$Attr_1, Attr_2 \dots Attr_k \rightarrow site\_yes$$

and

$$Attr_1, Attr_2 \dots Attr_k, Attr_{k+1} \dots \rightarrow site\_yes$$

we will take into account the second one. This means that the first rule is redundant; we have to eliminate it. On the other hand, it is a time consuming process to compute all the rules, and a faster solution that compute only a rules having no more that K attributes.

The algorithm 1 details our idea.

5.2 Validation Tests

In the aim to validate this algorithm we implement a cross validation test by leave one out. We implement the algorithm 1 and in a for loop we use it to predict for each specie in database its planting site. We counted the number of prediction errors and also the number of species with all planting sites that were well predicted.

Setting min support to 0.1 produces a huge numbers of rules with a weak real support and we can deduce very often both $site\_no$ and $site\_no$. So, this cross validation was run for a support set to 0.15 and varying the min confidence

We computed also a baseline, we randomly assign 'yes' or 'no' to each specie and each planting site according to the real frequency of the values 'yes' and 'no' for the corresponding planting site.

The table 2 of the results shows very good results comparing with the baseline.

Data: A collection $\mathcal{S}$ of species with known planting sites, a new specie T
Result: All possible planting sites for T, like attributes $site_i\_no$ or $site_i\_yes$
Compute $\mathcal{R}_T$ the set of restricted planting association rules;
Prune $\mathcal{R}_T$ into $\mathcal{AR}$ having not redundant rules;
for *every possible planting site p* **do**
  $Yes \leftarrow \{R \in \mathcal{AR} | rightside(R) = p\_yes\}$;
  $No \leftarrow \{R \in \mathcal{AR} | rightside(R) = p\_no\}$;
  $cumulated\_lift\_yes \leftarrow \sum_{R \in Yes} lift(R)$;
  $cumulated\_lift\_no \leftarrow \sum_{R \in No} lift(R)$;
  if $cumulated\_lift\_yes > cumulated\_lift\_no$ **then**
    fix p to *yes*;
  else
    fix p to *no*;
  end
end

Algorithm 1: Algorithm_predict_planting_sites.

Table 2: Results of the prediction algorithm applied the 134 trees database by leave-one-out cross validation.

Parameters	nb. errors	err. rate	correct
Baseline	235	0.292	11
sup=0.15, conf=0.9	149	0.18	50
sup=0.15, conf=0.95	151	0.19	50
sup=0.10, conf=0.9	154	0.2	52

6 CONCLUSION

In the present work we present some results obtained in the search for associations rules related to tree's planting sites. After exploring the important number of possible itemsets, the decision to eliminate low frequency items seemed to be an interesting strategy to generate rules. Although different rules were obtained for each planting sites, rules with the highest support and lift were explored and used to obtain the tree's genus. We observe that as in the study of items, genus Pinus and Quercus are the mostly frequent recommended for most of the planting sites, fortunately many other different genus are proposed strengthening tree's diversity.

We also use the generated association rules to predict the planting site. The obtained results should be verified by arborist experts. If positive, the method can be applied to other urban species collections. Algorithmically, the main concern was validation. For this validation procedure was implemented, showing low error rates for different parameter's combinations.

The main drawback of this prediction approach could be the execution time of the algorithm 1 and we will continue to work to reduce the time and the space complexities.

ACKNOWLEDGEMENTS

Héctor Javier Vázquez acknowledges the Institut H. Fayol, Ecole Nationale Suprieure des Mines, Saint Etienne, France for their invitation as a visiting researcher in 2015 and to the Universidad Autónoma Metropolitana (Azcapotzalco) for the permit to leave.

REFERENCES

Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD Record*, volume 22, pages 207–216. ACM.

Badii, M., Landeros, J., and Cerna, E. (2008). Patrones de asociación de especies y sustentabilidad (species association patterns and sustainibility). *Daena: International Journal of Good Conscience*, 3:632–660.

Hahsler, M. and Chelluboina, S. (2011). Visualizing association rules: Introduction to the r-extension package arulesviz. *R project module*, pages 223–238.

Hahsler, M., Gruen, B., and Hornik, K. (2005). arules – A computational environment for mining association rules and frequent item sets. *Journal of Statistical Software*, 14(15):1–25.

Kuhns, M. and Rupp, L. (2000). Selecting and planting landscape trees.

Lallich, S., Teytaud, O., and Prudhomme, E. (2007). Association rule interestingness: measure and statistical validation. In *Quality measures in data mining*, pages 251–275. Springer.

Longman, K. A. et al. (1993). *Tropical trees: propagation and planting manuals. Volume 1. Rooting cuttings of tropical trees.* Commonwealth Science Council.

SMA (2000). *Manual Técnico para la Poda, Derribo y Transplante de Árboles y Arbustos de la Ciudad de México, Secretaría del Medio Ambiente del Distrito Federal*. Secretaría del Medio Ambiente del Distrito Federal, México, D.F. http://www.sma.df.gob.mx/drupc/capacitacion/.

Vazquez, H. J. and Juganaru-Mathieu, M. (2014). Handling missing data in a tree species catalog proposed for reforesting mexico city. In *6th International Conference on Knowledge Discovery and Information Retrieval*, pages 457–464.

Watson, G. (2011). Fifteen years of urban tree planting and establishment research, trees, people and the built environment. In *Proceedings of the Urban Trees Research Conference*, pages 63–72.

Zaki, M. J. and Wagner Meira, J. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.