Customer Behavior Characterization Through Trajectory Reconstruction

Santiago Garcia Carbajal

University of Oviedo, Gijon , Asturias, Gijon, Spain.

E-mail: sgarcia@uniovi.es.com

David Corne

Heriot Watt University, Edinburgh, Scotland.

Summary. In this work we study the behavior of customers inside a big mall. To to that, we start by collecting GPS coordinates from customers during their visit to the store.

Keywords: Business Intelligence, Unsupervised Learning, Customer Characterization

1. State of The Art

Previous works ...

2. Our Case Study

In this section we describe the real case we are working on. The shop is a DIY store 50m x 50m, divided in logistic 8 sections:

- PLUMBING
- ELECTRICITY
- UNNAMED: generic section, dedicated to miscellanea
- WOOD
- PAINTINGS

- CERAMICS
- BUILDING
- UNDEFINED_AREA: corridors,showing rooms, etc

Our client has provided us with relevant data describing when and where a cell phone was detected inside the mall. Each time the

2.1. The Data

Data corresponding to customer behavior comes in this way:

```
198855222702,2017-04-06 08:46:04 UTC,26.0,34.0,1
198855222702,2017-04-06 08:48:00 UTC,30.0,36.0,1
198855222702,2017-04-06 08:49:03 UTC,40.0,34.0,1
198855222702,2017-04-06 08:51:16 UTC,32.0,20.0,1
198855222702,2017-04-06 08:53:18 UTC,33.0,34.0,1
198855222702,2017-04-06 08:59:19 UTC,41.0,32.0,1
198855222702,2017-04-06 09:03:22 UTC,38.0,38.0,1
198855972730,2017-04-03 05:18:40 UTC,41.0,33.0,0
198855972730,2017-04-03 05:19:13 UTC,16.0,38.0,0
198855972730,2017-04-03 05:20:04 UTC,22.0,37.0,0
198855972730,2017-04-03 05:22:05 UTC,16.0,40.0,0
198855972730,2017-04-03 05:23:57 UTC,39.0,28.0,0
198855972730,2017-04-03 05:32:10 UTC,26.0,33.0,0
198855972730,2017-04-03 05:33:28 UTC,3.0,20.0,0
```

This is, a huge file containing MAC, UTC, and a sequence number indicating if that MAC has been inside the shop before, and how many times. In the example, we can see acustomer detected seven times, and another customer detected eight times.

Initially, we could translate this coordinates into a trajectory. But, shops are full of lineals and, organized in corridors and areas. So, we need to guess how the real trajectory was starting from this file, and knowing where the obstacles are. To do that, we use

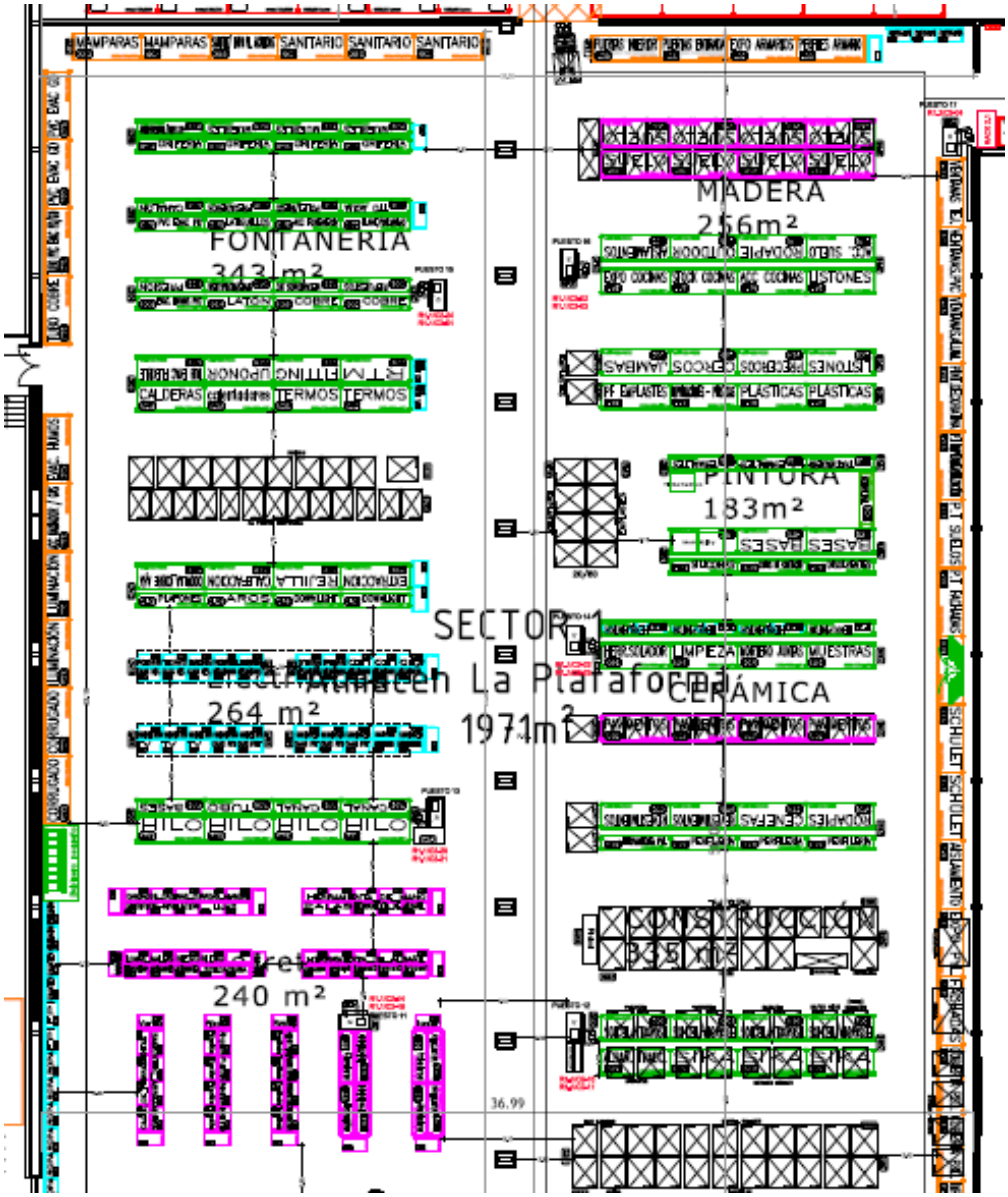

Fig. 1. The Shop.

Lee algorithm. In figure 2, the assumed behavior of a customer is shown. Red squares represent the points where the cell phone was detected. Yellow squares are generated by Lee Algorithm, avoiding obstacles (Grey squares).

From this file, assuming that the trajectory built by Lee Algorithm is realistic enough, and through simple calculations, we can also extract some numeric features of any trajectory. Some of them are:

Entering Time Directly extracted from file (UTC)

Leaving Time Same

Staying Time Calculated as the difference, in seconds, between Entering and Leaving Time

Total Path Lenght Calculated as the number of yellow and red squares genereted by Lee Algorithm

Average Speed Total Path Length divided by Staying Time

Detection Points Number of red squares

Redundancy Percentage of times that the customer steps on the same square

These magnitudes will be used as basic inputs for the unsupervised clustering algorithm that we will be using on the data. But apart from them, we will be using also second order values that we calculate studying the trajectories, in terms of the sequence and fraction of different logistic areas that bthe customer is visiting. In order to do that, we work with a logical description of the shop, in terms of labels naming different areas dedicated to each kind of product.

3. Clustering

Cluster Analysis is one of the vast groups of Data Mining Techniques. The main idea is to segment a customer database so that customers within segments are similar, and different from customers in other segments. Similarity is measured in terms of the "Clustering Variables", which may be psychographics, demographics, or transaction measures such

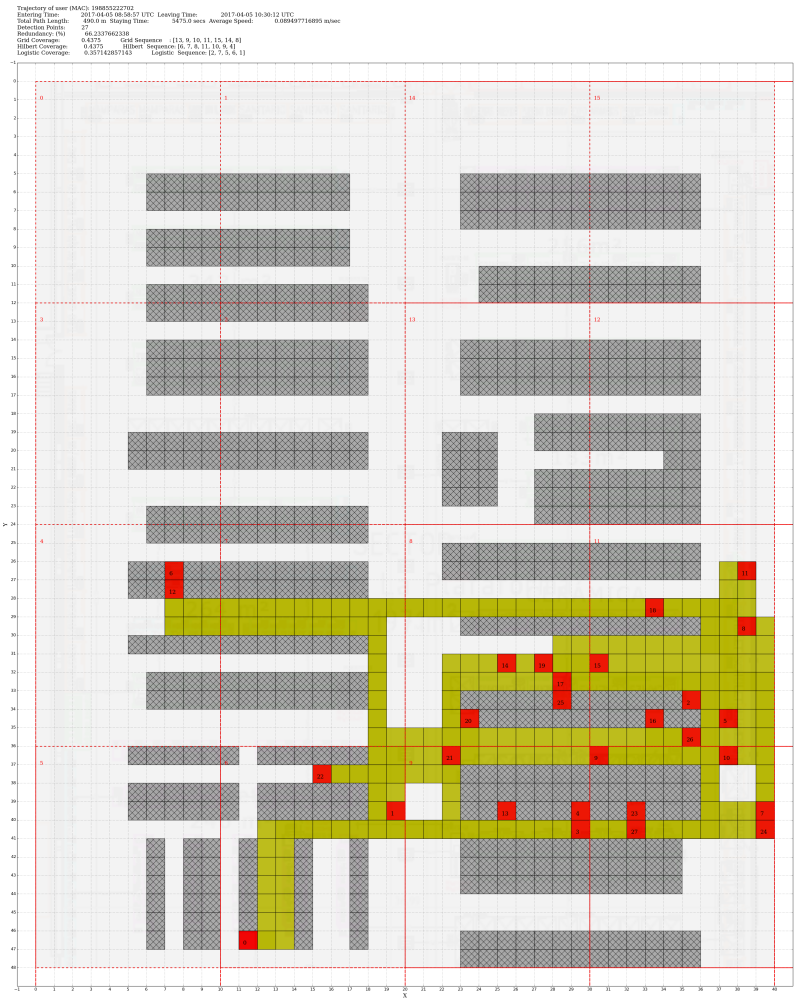

Fig. 2. Reconstructed Trajectory.

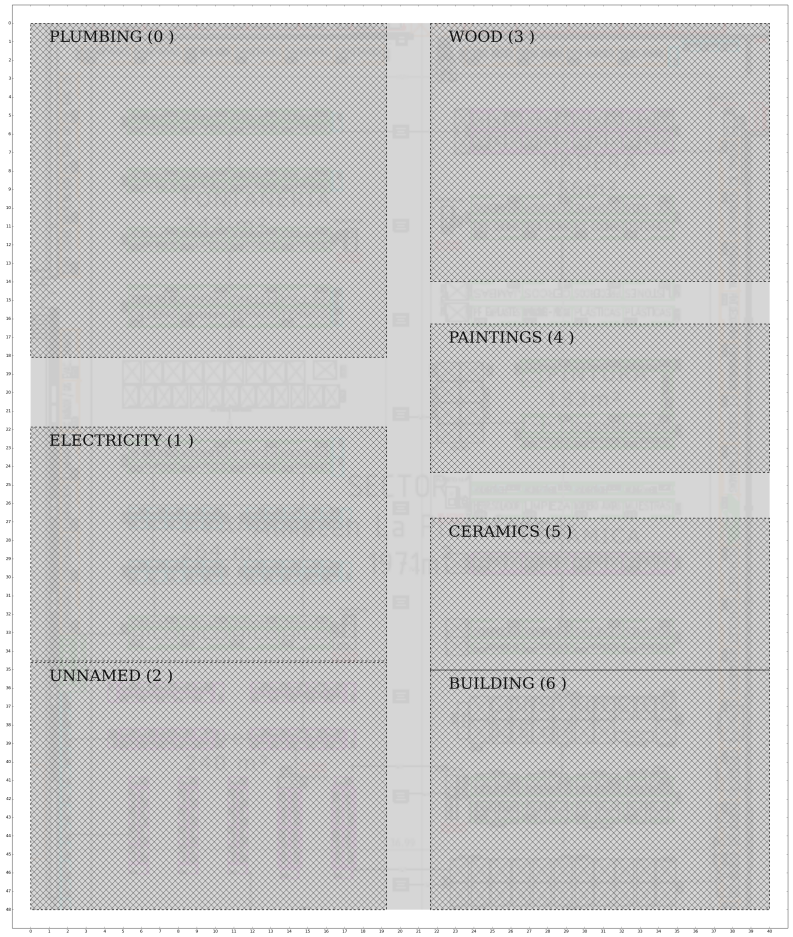

Fig. 3. The Shop.

as recency, frequency or monetary value. The clusters allow rich interpretation with strong implications for which customers should be targeted with a particular offer or marketed to in a certain way. Unlike the classification process, where a class label is given for each customer, and based on this label elements are classified. In this case, the class label of each customer is unknown. Through clustering analysis, these groupings are discovered. Clustering is the process of partitioning a set of data objects into subsets. Each subset is a cluster, and objects in a cluster are similar to one another, yet dissimilar to objects in other clusters. In this context, different clustering methods may generate different clusterings on the same data set. The partitioning is not performed by people, but by the clustering algorithm. Consequently, clustering is useful in that it can lead to the discovery of previously unknown groups within the data. Cluster analysis has been widely used in many applications such as image pattern recognition , business intelligence , information retrieval, biology and security. In business intelligence, clustering can be used to organize a large number of customers into groups, where customers within a group share strong, similar characteristics.

3.1. Clustering Process

To conduct a cluster analysis, several steps are necessary:

(a) Select variables on which to cluster
(b) Select a similarity measure and scale the variables
(c) Select a clustering method
(d) Determine the number of clusters
(e) Conduct the cluster analysis, interpret the results, and apply them

3.1.1. Selection of variables and normalization

...

3.1.2. K-Means Algorithm

The Kmeans may be the most popular clustering method among data miners. This algorithm is a centroidbased technique and belongs to the family of partitioning methods.

Let us explain how this algorithm works: Suppose a data set D contains n objects in Euclidean space. Partitioning methods distribute the objects in D into k clusters, C1; : : : ;Ck , that is, Ci D and Ci Cj = ? for (1 i; j k). An objective function is used to assess the partitioning quality so that objects within a cluster are similar to one another, but dissimilar to objects in other clusters. A centroid based partitioning technique uses the centroid of a cluster, Ci, to represent that cluster. Conceptually, the centroid of a cluster is its center point. The centroid can be defined in various ways such as by the mean or medoid of the objects (or points) assigned to the cluster. The difference between an object p 2 Ci and ci, the representative of the cluster, is measured by dist(p; ci), where dist(x; y) is the Euclidean distance between two points x and y. The quality of the cluster Ci can be measured by the withincluster variation, which is the sum of squared error between all objects in Ci and the centroid ci defined as:

FORMULA

where E is the sum of the squared error for all objects in the data set; p is the point in space representing a given object; and ci is the centroid of cluster Ci. This objective function tries to make the resulting k clusters as compact and as separate as possible. Optimizing the withincluster variation is computationally challenging. In the worst case, we would have to enumerate a number of possible partitions that are exponential to the number of clusters, and check the withincluster variation values. It has been shown that the problem is NPhard in general Euclidean space even for two clusters. Moreover, the problem is NPhard for a general number of clusters k even in the 2D Euclidean space. If the number of clusters k and the dimensionality of the space d are xed, the problem can be solved in time O(ndk+1 log n) where n is the number of objects. This algorithm denes the centroid of a cluster as the mean value of the points within the cluster. It proceeds as follows. First, it randomly selects k of the objects in D, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is most similar, based on the Euclidean distance between the object and the cluster mean. The kmeans algorithm then iteratively improves the withincluster variation. For each cluster, it computes the new mean using the objects assigned to the cluster in the previous iteration. All the objects are then reassigned using the updated

means as the new cluster centers. The iterations continue until the assignment is stable, that is, the clusters formed in the current round are the same size as those formed in the previous round. The kmeans algorithm is described in algorithm 1.

3.1.3. Automatically Determining the Number of Clusters

Determining the appropriate number of clusters is one of the most difficult problems in clustering. Usually, the criteria chosen tends to be subjective. For example, the relative sizes of the clusters should be large enough to be managerially meaningful. The clusters with few elements (customers) may be treated as outliers and ignored. The methods for determining the number of clusters depends on the clustering algorithm being used and there are several criteria available. However, Milligan [80] showed that the procedure by Calinski [21] performed the best among 30 different criteria. The following criterion suggested in [21] is used to determine the number of clusters:

FORMULA

G(k) = (n k)(T W) (k 1)W (2.8) where k is the number of clusters, n is the number of customers, W is the square sum of the distances of the customers to the center of its cluster, and T is the square sum of the dierences of each customer to the average customer, essentially, the center of the full data. The optimal number of clusters can be determined by selecting k which returns the maximum value for G(k), because in that case, W, or the distances between customers and the center of their clusters, is relatively small compared to T, the distances between customers and the center of the entire data.

3.2. How to Leave Comments

Comments can be added to the margins of the document using the todo command, as shown in the example on the right. You can also add inline comments:

Here's a comment in the margin!

This is an inline comment.

3.3. How to Make Tables

Use the table and tabular commands for basic tables — see Table 1, for example.

Table 1. An example table.

Item	Quantity
Widgets	42
Gadgets	13

3.4. How to Write Mathematics

LaTeX is great at typesetting mathematics. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

3.5. How to Make Sections and Subsections

Use section and subsection commands to organize your document. LaTeX handles all the formatting and numbering automatically. Use ref and label commands for cross-references.

3.6. How to Make Lists

You can make lists with automatic numbering ...

(a) Like this,
(b) and like this.

... or bullet points ...

- Like this,
- and like this.

... or with words and descriptions ...

Word Definition

Concept Explanation

Idea Text

4. Results

4.1. Class 0

4.2. Class 1

4.3. Class 2

4.4. Class 3

5. Citations and References

Here are some natbib examples. You can cite examples using the citation key (Trang and Mebkhout, 1983) in your .bib file. (On Overleaf, you can access the .bib file via the Project menu.) There are commands for in-text citations, like Goresky and MacPherson (1981). And you can pass an option to specify additional details, such as a page or chapter number, as an option (Fulton, 1983, p. 130).

We hope you find Overleaf useful, and please let us know if you have any feedback using the help menu above.

References

Fulton, W. (1983) Introduction to intersection theory in algebraic geometry. In *Regional Conference Series in Mathematics*, no. 54.

Goresky, M. and MacPherson, R. (1981) On the topology of complex algebraic maps. In *Algebraic Geometry Proceedings, La Rábida, Lecture Notes in Mathematics*, no. 961.

Trang, L. D. and Mebkhout, Z. (1983) Variétés caractéristiques et variétés polaires. *C. R. Acad. Sc. Paris*, **296**, 129–132.

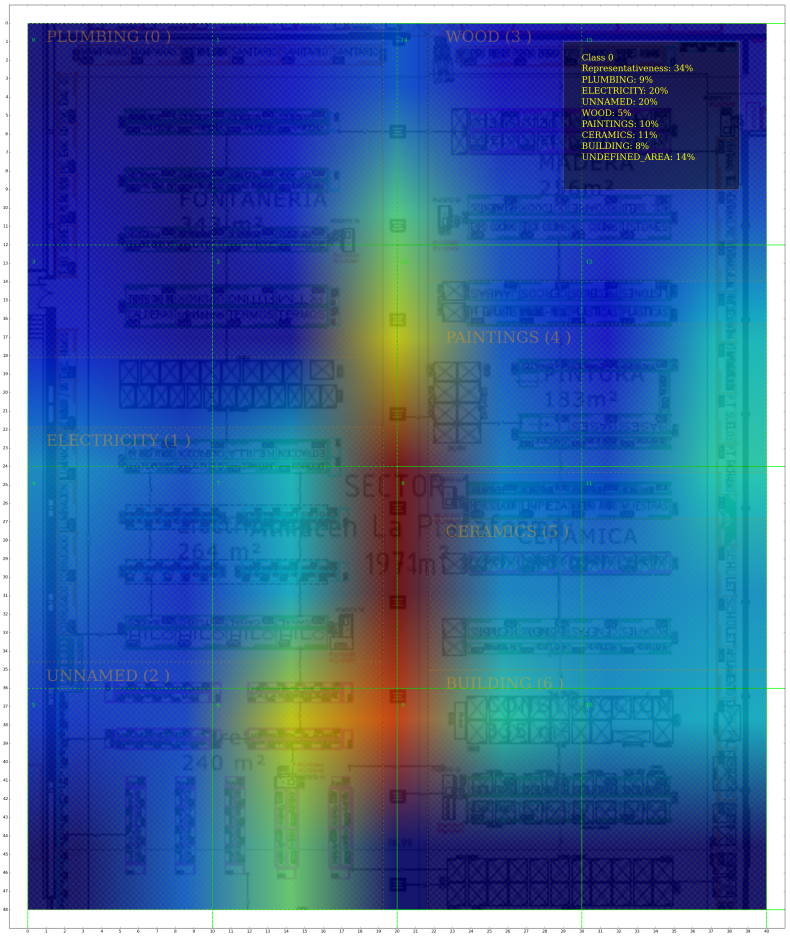

Fig. 4. Class0 Heatmap as determined by kMeans Algorithm .

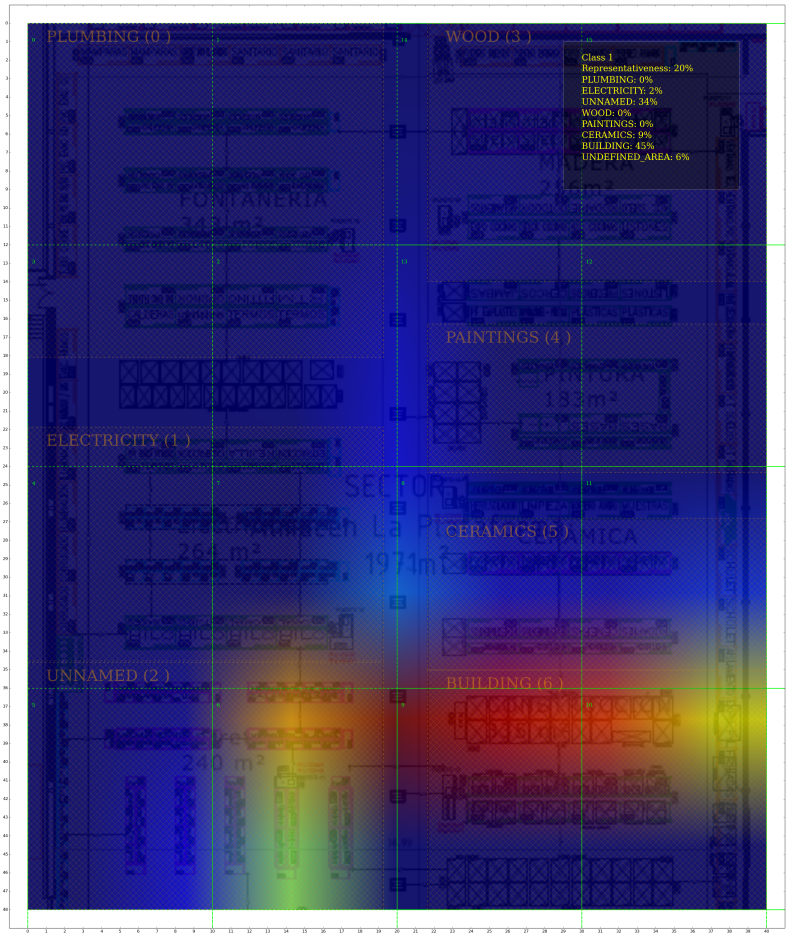

Fig. 5. Class1 Heatmap as determined by kMeans Algorithm .

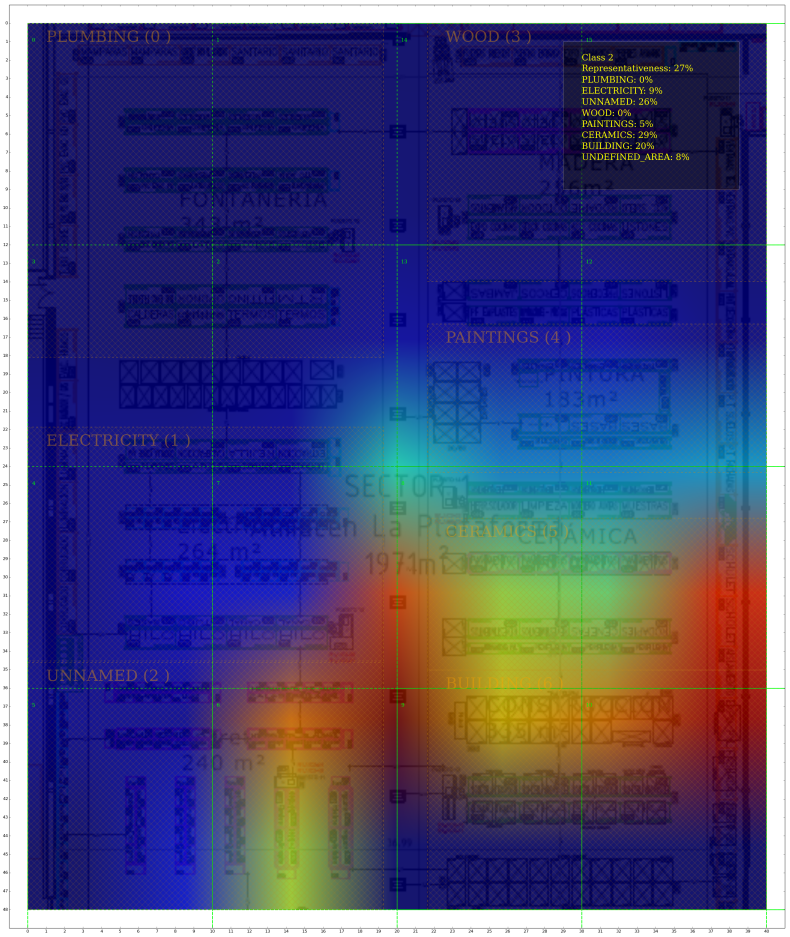

Fig. 6. Class2 Heatmap as determined by kMeans Algorithm .

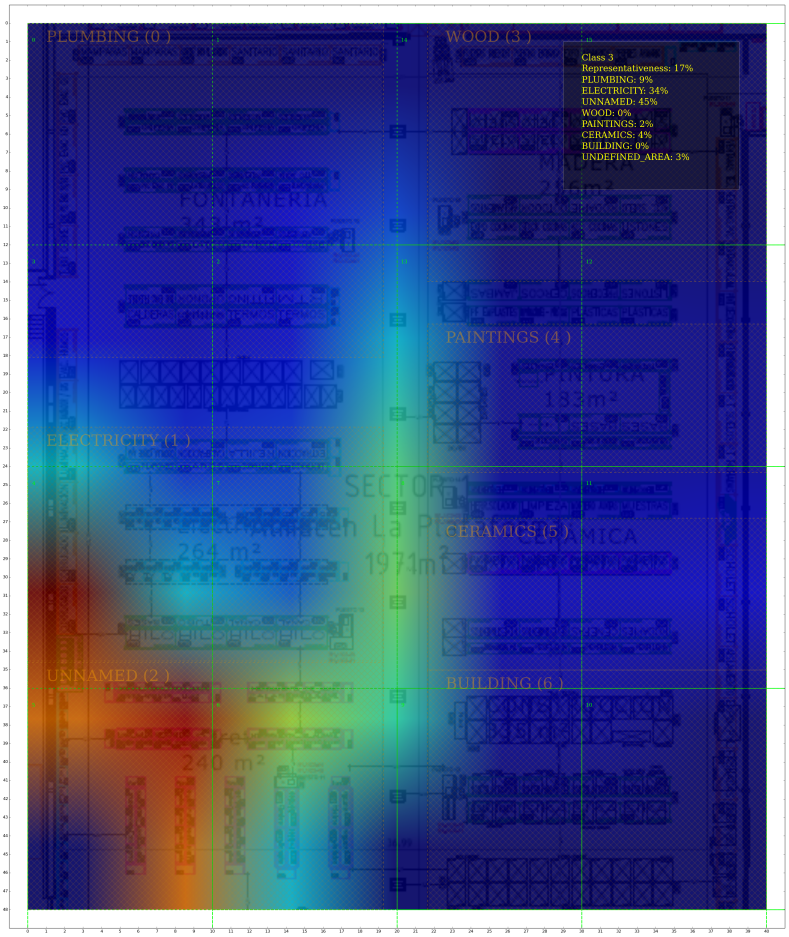

Fig. 7. Class3 Heatmap as determined by kMeans Algorithm .