**MASTER TRIED**
NELSON FERNANDEZ-PINTO
CARL ROBINSON
15th January 2018

# Etude de Cas. Part I

## Objectives

The "Exploitation of the OQAI Housing Survey" project provided an analysis of determinants of indoor air quality in dwellings, and in particular the toxic pollutant formaldehyde. Our objective is perform an initial data exploration of project data, with the aim of discovering a research question that warrants further study.

## Data and methods

The file data_hab_men_log_Pertinentes.xls contains data on 567 houses surveyed, across the 30 variables found to be the most pertinent in the previous project. These describe features of the houses themselves, environment inside the house, and various details about the residents and their living habits. This file contained mostly qualitative variables, but also a small number of quantitative variables which needed to be transformed into categorical variables (e.g. low/medium/high) for ease of analysis.
The file Polluant-5.txt contains numerical measurement data of 14 pollutants in the 567 houses surveyed. These too were all transformed into categorical variables using quartiles (e.g. low/medium/high) for analysis.
All samples containing empty or NaN values for any variable were removed completely, and the remaining 530 samples of categorical data from both files were merged together into a single table.
In total there are 44 variables of which 23 are numerical. Of these, there are 14 toxic chemicals, 7 behavioural variables, plus age and revenue of residents.) The rest of the variables are categorical, of type integer.
Various data analysis methods were employed, including multivariate correspondence analysis, pearson correlation coefficients and matrices, and fully connected neural networks.

## I. Data exploration

A number of pertinent quantitative behavioural variables relating to use of chemical products in the household exist in the data set, and were chosen for initial consideration, in relation to concentrations of Formaldehyde and other chemicals of the same family

| Quantitative Variable | Description |
|---|---|
| Formaldehyde / Acetaldehyde / Hexaldehyde | Float value indicating average concentration of each chemical |
| Deodorants.ICOS1 | Float value indicating relative usage of deodorant |
| EauDeToilette.ICOS2 | Float value indicating relative usage of eau de toilette |
| ProdSoinCheveux.ICOS3 | Float value indicating relative usage of hair care products |
| ProdSoinVisage.ICOS4 | Float value indicating relative usage of facial skincare products |
| VernisOngleDissolv.ICOS5 | Float value indicating relative usage of nail polish remover |
| DesodoAutreEnsens.QPD2b | Float value indicating average number of times per week another type of deodoriser is used (e.g. candle, lamp, pot pourri, bloc WC) |
| ACTIVITE | Float value indicating <?> |
| Age | Integer indicating the age of house owner |

Table 1. Table of proposed quantitative variables

## Principal Component Analysis

To discover relationships between these quantitative variables and the measured air pollutants, a principal component analysis (PCA) was carried out. In order to simplify the resulting visualisation and its analysis, we considered only the aldehyde family (formaldehyde, hexaldehyde et acetaldehyde).

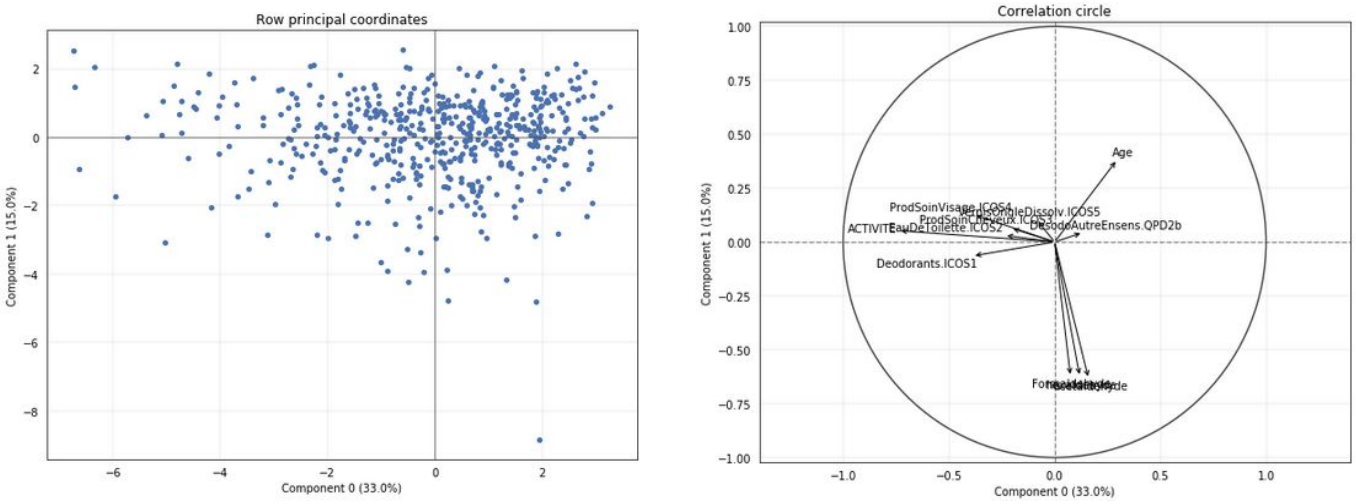
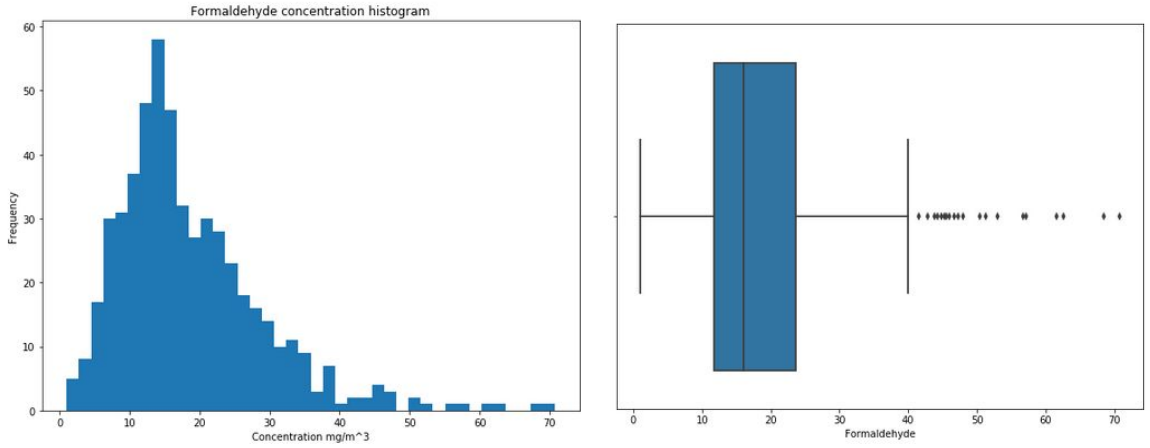



Figure 1. PCA of quantitative behavioural variables and Formaldehyde family

- In the correlation circle we observe three principal groups of correlated variables: the formaldehyde family, age, and behavioural features (use of personal care products and activity).
- It is clear that all chemicals in the aldehyde family are highly correlated i.e. they are often present and increase together, as seen by the arrows in the correlation circle all pointing in the same direction, with the same length. The point almost directly downwards, contributing the vast majority of the variance in the second principal axis. This accounts for the vertical spread in the point cloud.
- To eliminate redundancy and reduce the number of variables we are working with, we chose to consider only one of these: formaldehyde, a dangerous carcinogen[1], which we took to represent the whole group.
- The behavioural variables appear to be somewhat correlated with each other, however the arrows are shorter in length, and the angles are more diverse. Also, each one contributes little to the overall variance along the first principal axis, showing that none is particularly dominant (the Activite variable contributes the most).
- Most important, the behavioural variables appear to be poorly correlated to the aldehyde group, and other being roughly anti-correlated. However, there might be non-linear correlations that are not possible to reveal with this linear transformation method.
- Given this result, and the low total inertia of the representation (48%), we hypothesise that non-linear correlations exist. Then, the use of non-linear techniques such as t-distributed Stochastic Neighbor Embedding (t-SNE) and rank correlations could be more effective as revealing the relationships in the data.

## Univariate analysis. Formaldehyde concentration

As previously stated, high concentrations of formaldehyde are toxic to human ADN, and can degenerate in chronic diseases as cancer and respiratory insufficiency. According to the National Center of Biotechnology (NCBI) guidelines, health risks depend on the concentration and time of exposure, being the limit of 0.1mg/m$^3$ considered as safe to prevent health problems.[2]. The following histogram and box plot, shows the centration of Formaldehyde (in *ug/m*$^3$) across the 530 housings studied.




[1] C. Bosetti, J. K. McLaughlin, R. E. Tarone, E. Pira, C. La Vecchia; Formaldehyde and cancer risk: a quantitative review of cohort studies through 2006, *Annals of Oncology*, Volume 19, Issue 1, 1 January 2008, Pages 29–43, https://doi.org/10.1093/annonc/mdm202

[2] NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health. WHO Guidelines for Indoor Air Quality: Selected Pollutants. Geneva: World Health Organization; 2010.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **Formaldehyde** | 530.0 | 18.603283 | 10.483351 | 1.02 | 11.705 | 16.005 | 23.6325 | 70.75 |

Figure 2. Histogram and boxplot of Formaldehyde concentrations.

- Formaldehyde levels across the houses surveyed have a rough gaussian distribution, slightly skewed to the right in the higher concentrations of the pollutant. The average concentration is 18.60ug/$m^3$ and 75% of the housings are lower than 23.63ug/$m^3$ .
- The box plot reveals that the skewness is due to the existence of several outliers; measurements from 40 to 70 ug/$m^3$. Although these measures are relatively high, still remains under the safety guidelines.
- The mean of the distribution is 18.6ug/$m^3$ however is affected by outliers, the median (16.8ug/$m^3$) can be a better estimator of the real center of the distribution across the population.

Based on the quartiles discovered during the previous analysis we propose the following ranges to categorize the concentration of formaldehyde:

| Category | Concentration ug/$m^3$ |
|---|---|
| Low | 1 - 11.70 ug/$m^3$ |
| Intermediate | 11.70 - 23.63 ug/$m^3$ |
| High | 23.63 - 40 ug/$m^3$ |
| Very high | > 40 ug/$m^3$ |

Table 2. Table of Formaldehyde concentration ranges.

The aforementioned quantization allowed to obtain a repartition that preserves the original distribution. Most of samples are in the “intermediate” box, low and high classes are almost equally represented. The category “very high” is basically composed by outliers.

| | Formaldehyde | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | count | mean | std | min | 25% | 50% | 75% | max |
| class | | | | | | | | |
| **high** | 113.0 | 29.523186 | 4.392178 | 23.65 | 25.8000 | 28.560 | 32.7800 | 39.96 |
| **intermediate** | 264.0 | 16.739129 | 3.444187 | 11.75 | 13.7975 | 16.005 | 19.8325 | 23.58 |
| **low** | 133.0 | 8.093534 | 2.501351 | 1.02 | 6.7000 | 8.220 | 10.1700 | 11.69 |
| **very high** | 20.0 | 51.402500 | 8.684002 | 41.60 | 45.1750 | 47.645 | 56.8350 | 70.75 |

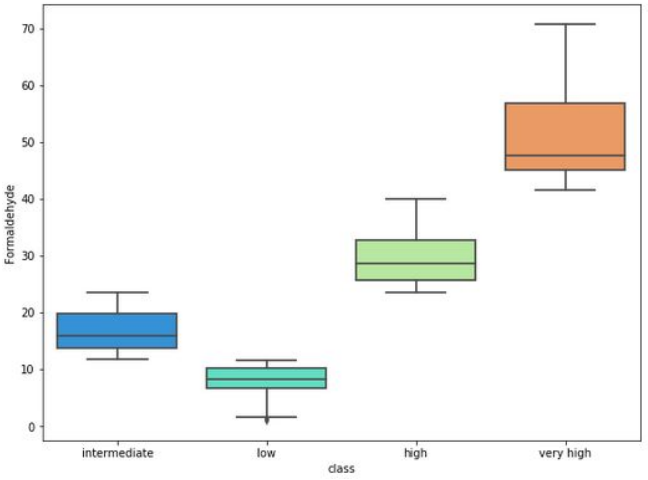



Figure 3. Table and boxplot of formaldehyde ranges after quantization.

## t-SNE visualization

We now employ the t-distributed Stochastic Neighbor Embedding (t-SNE) analysis method. t-SNE is able to provide a 2D visual representation of high-dimensional data that preserves the original structure. The method performs a non-linear transformation on the data, which allows us to identify non-linear relationships that PCA is unable to expose. As t-SNE only works with numerical variables, we used as input all quantitative behavioural variables plus the formaldehyde concentration (i.e. all other chemicals were omitted). In figure 4, the coloured categories are those created in the previous step.

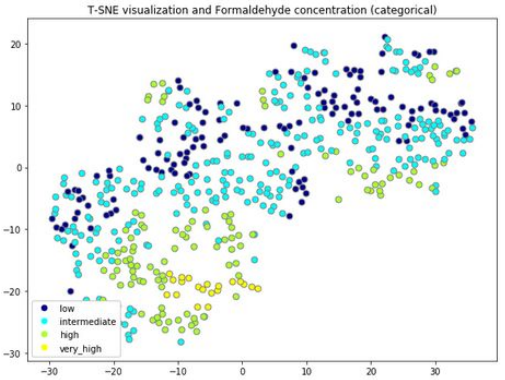



Figure 4. t-SNE clustering

- We can now see some homogenous clusters start to appear. The high (green) and very high (yellow) values are always grouped together, as are the low (blue) and intermediate (cyan) concentrations.
- We observe a cluster of housings with high and very high formaldehyde concentrations located in the lower part of the plan. Moderate and low concentration are widely spread.
- There are multiple clusters for each category, and the clusters that do exist are not separated by any empty space, suggesting manifold entanglement.
- This may mean that the data in its raw form is not clearly separable into clusters. Another explanation could be because the Manifold assumption cannot be true for any arbitrary manifold dimension, and so in our case a 2D representation may simply insufficient to reveal the clusters that exist.

## Spearman's correlations using rank ordered variables

Spearman's correlation, also known as rank correlation, measures the relationship between rankings of different ordinal variables or different rankings of the same variable, where a "ranking" is the assignment of the ordering labels "first", "second", "third", etc. to different observations of a particular variable. A rank correlation coefficient measures the degree of similarity between two rankings, and can be used to assess the significance of the relation between them. Then, spearman's correlation allows to identify if two variables increases monotonically in a rank without a normal distribution assumption. All quantitative and qualitative behavioural variables were converted to integer rank values based on their relative magnitude as above, then correlated with **formaldehyde** concentrations.

The table of correlations shows that the following variables are somewhat correlated:

| | correlation | pvalue | variable |
|---|---|---|---|
| **26** | 0.251468 | 4.354151e-09 | NIACe1 |
| **17** | 0.172533 | 6.534718e-05 | Nb_Enfants_inf_10 |
| **16** | 0.122421 | 4.767889e-03 | Nb_Person |
| **25** | 0.118988 | 6.095682e-03 | HSRF |
| **0** | 0.116827 | 7.093487e-03 | Deodorants.ICOS1 |
| **24** | 0.091049 | 3.612659e-02 | Structure_menage |

- **Age of construction:** This is the most correlated variable, with a significant coefficient value of 0.25, suggesting more recent housings present higher concentrations. That is consistent with the fact that formaldehyde concentrations decreases with time[1].
- **Housing variables:** Number of children and people living in the house, dimension and type of structure, allows to characterize the type of housing that is generally most susceptible to present higher formaldehyde concentrations.
- **Deodorants**: Air fresheners and other escences are reported to have high formaldehyde concentrations[3].
- In general the correlations are not high. However, due to the statistically representative population and the complexity of the data set, they reflects relationships that actually exist between variables.

## Spearman's correlations using one-hot encoding

One-hot encoding is a technique that allows to represent all the possible states of a categorical variable in a binary representation. Each ordered range of a variable is therefore converted to a new auxiliary variable. According to our line of study, formaldehyde concentration is converted to four mutually exclusive binary variables: formaldehyde_low, formaldeyde_intermediate, formaldehyde_high and formaldehyde_very high.

This approach allows to correlate each modality separately, revealing relationships that exist specifically in a range of interest. In our case, we focus our attention on very high and high formaldehyde concentrations. The following table shows the higher correlations obtained.

This approach reveals previously unseen possible correlations:

| | Formaldehyde_very high |
|---|---|
| **Formaldehyde_very high** | 1.000000 |
| **Nb_Person_4** | 0.125147 |
| **VoitureDansGarage.DGG3n_2** | 0.090867 |
| **Jardinage.TMG7n_4** | 0.089166 |
| **VoitureDansGarage.DGG3n_1** | 0.085515 |
| **Occupation_1** | 0.085044 |
| **TraitementParasitesAnimaux.ANTCPn_3** | 0.085035 |
| **Age_2** | 0.082594 |
| **Jardinage.TMG7n_5** | 0.074513 |

- **Vehicles emissions:** People that leave started vehicles waiting inside the garage on daily basis. That is consistent with the fact that motor vehicles emit several polluants including formaldehyde[4].
- **Gardening:** Very frequent gardening activities. Some studies have confirmed that high traces of formaldehyde can be found on freshly cut grass[5].
- **Animal parasite treatment:** Formaldehyde is often used as biocide in parasite treatment of domestic animals[6]. Therefore it is present in common use medicaments, accesibles to all public.

[3] Royal College of Physicians - Every breath you take. The lifelong impact of air pollution. February 2016. https://www.rcplondon.ac.uk/projects/outputs/every-breath-we-take-lifelong-impact-air-pollution

[4] Air Toxics from Vehicles and Their Fuels https://www.uvm.edu/~empact/air/EPA_factsheet.html

[5] Air pollution and the smell of freshly cut grass. Wayne Kirstine, Ian Galbally and Martin Hooper. School of Applied Sciences, Monash University, Churchill, Vic 3842, Australia 2 CSIRO Division of Atmospheric Research, Aspendale, Vic 3195, Australia

[6] https://www.amazon.co.uk/Fish-Pond-Anti-Parasite-Treatment/dp/B010M774CW http://www.swelluk.com/nt-labs-formaldehyde/

As seen in the previous experience, correlation coefficients values are not especially high, ranging from |0.07| to |0.25|. However, given that we have a respectably large population (more than 500 examples) these values do reliably represent a relationship between the variables. No single variable is a reliable predictor of formaldehyde concentrations though, so multiple variables will need to be used to train any prediction model.

## Multiple Correspondence Analysis (MCA)

Multiple correspondence analysis (MCA) is a data analysis technique for categorical data, that represents data as points in a low-dimensional Euclidean space. To use MCA on our data, all numerical variables were first converted to categorical as follows:

| Variable | Numerical ranges of categories |
|---|---|
| Age of resident | Under 25 years, 25 to 45 years, 45 to 65 years, over 65s |
| HSRF | Small = below lower quartile, Medium = above lower quartile and below upper quartile, Large = above upper quartile |
| Age of property | Based on quartiles: Before 1948, 1948 to 1967, 1967 to 1989, after 1989 |
| Activity | Low = below lower quartile, Intermediate = above lower quartile and below upper quartile, High = above upper quartile |
| Revenue | Based on quartiles: Low = <1299, Moderate = 1299 to 2350, High = 2350 to 4300, Very High = >4300 |
| Deodorants.ICOS1 | Low = below lower quartile, Intermediate = above lower quartile and below mean, Intermediate = above mean and below upper quartile, Very High = above upper quartile |
| EauDeToilette.ICOS2 | Same as above |
| ProdSoinCheveux.ICOS3 | Same as above |
| ProdSoinVisage.ICOS4 | Same as above |
| VernisOngleDissolv.ICOS5 | Same as above |
| DesodoAutreEnsens.QPD2b | Same as above |

Table 3. Table of quantitative variables ranges.

First we project the instances onto a lower dimensional plane, and colour code the instances based on their formaldehyde concentration categorisation. The following figure shows the obtained MCA.

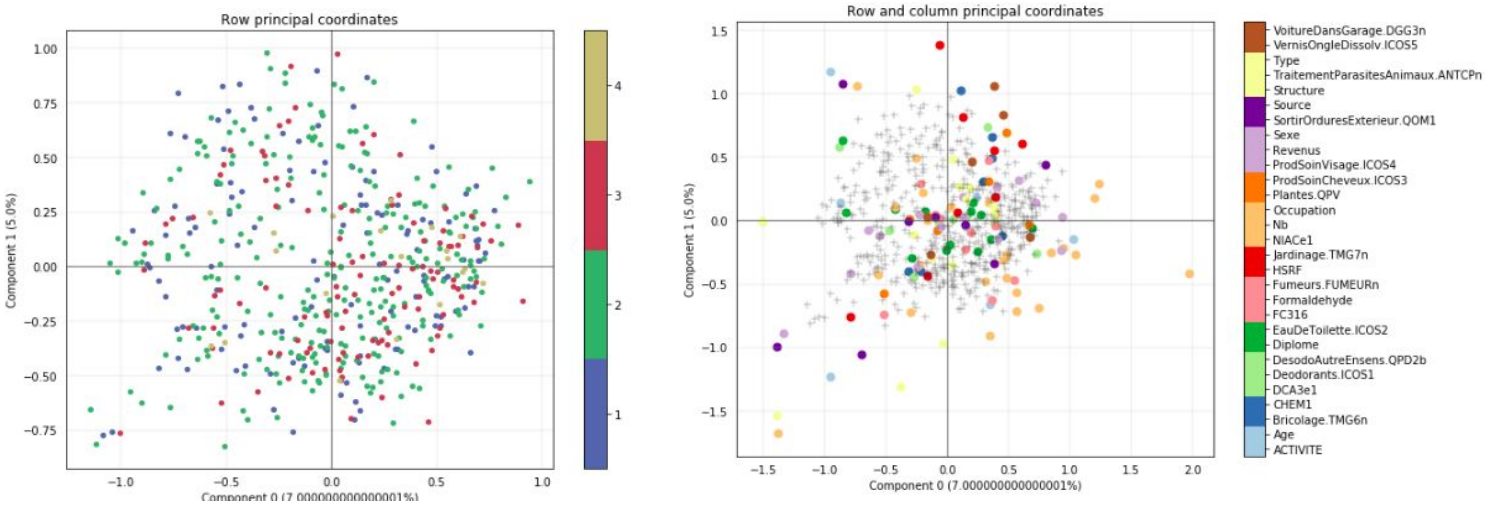



Figure 5. MCA and principal coordinates (row and column)

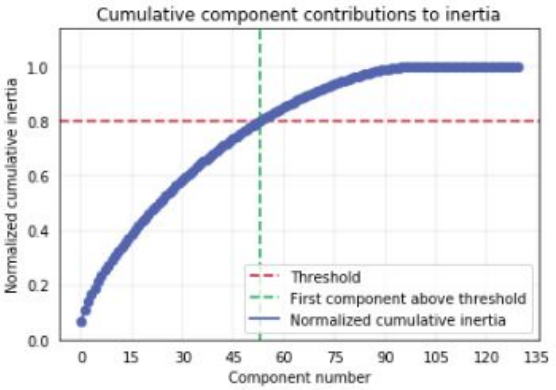
We can see that housing with low and intermediate concentrations is evenly distributed across the 2D plane, whereas high and very high concentrations have a tendency to be cluster towards the lower right. On its own, this tells us very little about the relationships in the data, however it will help us to interpret the variables projection.

Next we project the variables onto the 2D plane, to identify relationships between them. The method also helpfully clusters variables that have a high correlation coefficient together using the same colour.

We can see that in the bottom right corner there is a cluster of orange variables, corresponding to age of housing, number of people and occupation. As the examples and variables projections share the same orientation, we can conclude that that the high and very high concentrations of formaldehyde are correlated to high values for the aforementioned variables. This confirms our earlier findings using spearman's correlation.

This representation also reveals a previously unseen correlation, formaldehyde and cigar (colored in pink). Studies confirm that small concentrations of formaldehyde can be found in cigar smoke (3.4-8.4 ug/cigarette) and this effect becomes more important in poorly ventilated places[7].

The cumulative contributions curve, shows that more than 45 components are required in order to reach the threshold of 80% of the inertia. This shows that each variable contributes very little to the total variance in the dataset, and therefore the explanation of formaldehyde concentration.




## Agglomerative Hierarchical Clustering (CAH)

Agglomerative hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters with a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy. Here we experimented with different distance metrics (euclidean, hamming, cityblock, cosine) and with different linkage methods (single, complete, average, ward), to find the optimal clustering. We measured the clustering effectiveness using the Cophenetic Correlation Coefficient, which compares (correlates) the actual pairwise distances of all the samples to those implied by the hierarchical clustering. The closer the value is to 1, the better the clustering preserves the original distances. Our highest coefficient was 0.5627, obtained using a euclidean distance with average linkage, but this produced a very poor clustering. We then used euclidean distance with ward linkage, which had a lower coefficient of 0.4850, and this clustered the data very well, with large distances separating them in the dendrogram.

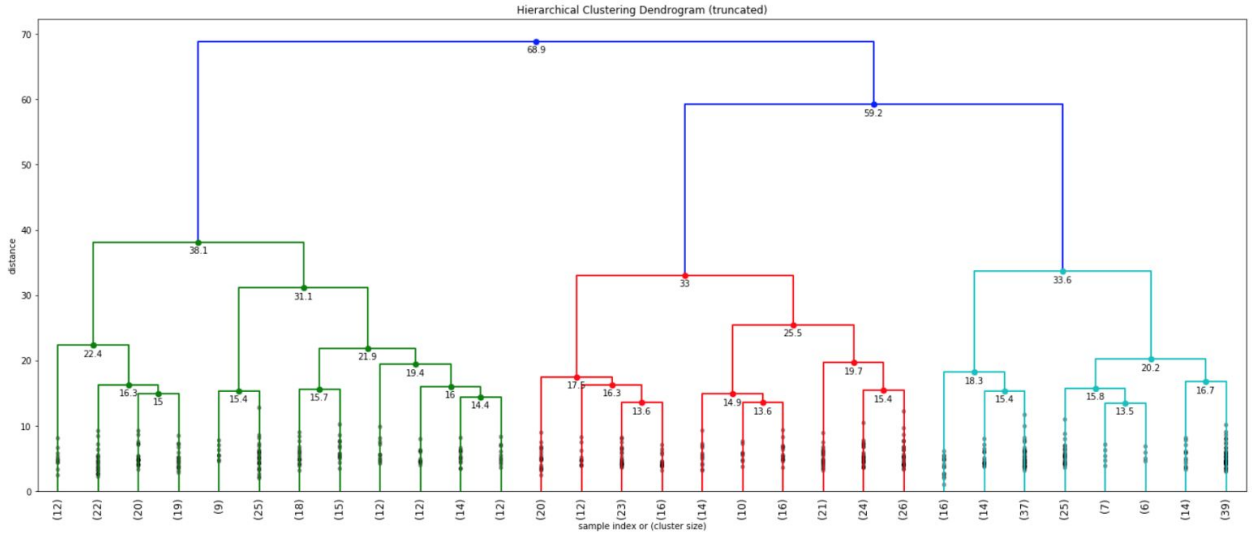
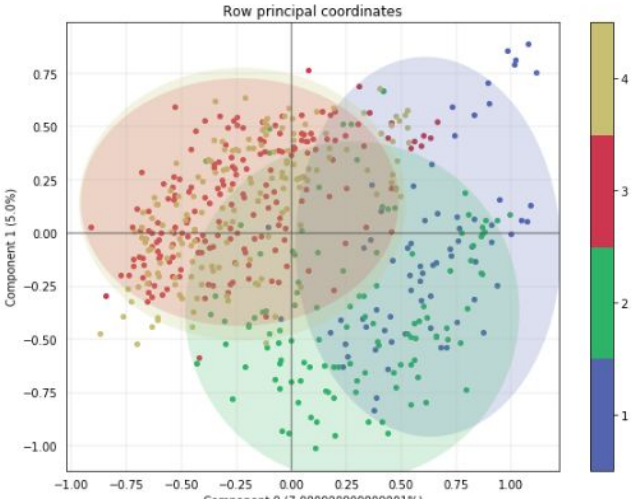



Figure 7. CAH dendrogram clustering (above) and MCA plot (below)

We chose to cut the dendrogram at K=4, and used the cluster assignments at this level to colour our MCA plot. The 4 clusters are shown by the ellipses, which are clearly separated with only some overlap. Clusters 3 and 4 actually overlap almost perfectly, and can therefore be merged into a single cluster in later analysis. The next part of this study will seek to characterise the houses found in each of these clusters, and explore the levels of formaldehyde.




## Conclusion

The exploratory data analysis allowed us to understand and characterise several factors that are associated to the increase on formaldehyde concentration. Variables as type and year of construction, use of deodorants, smoking, gardening and vehicle's emissions appear as possible causes of high formaldehyde concentrations. However, correlations found where low, indicating there is not an small set of variables that can fully explain formaldehyde data. Then, the information is distributed along the variables. As next step on this study, we propose to build a model to predict the range of formaldehyde concentration using a parametric supervised method: random forest.

[7] Formaldehyde determination in tobacco smoke--studies under experimental and actual conditions. Schaller KH1, Triebig G, Bever B. 1989.