Classification of dwellings into profiles regarding indoor air quality, and identification of indoor air pollution determinant factors

Jean-Baptiste Masson, Gérard Govaert, Corinne Mandin, Séverine Kirchner, André Cicollela

HAL Id: hal-00447841

https://hal.archives-ouvertes.fr/hal-00447841

Submitted on 16 Jan 2010

Classification of dwellings into profiles regarding indoor air quality, and identification of indoor air pollution determinant factors

Jean-Baptiste Masson[1,2] [*], Gérard Govaert[2], Corinne Mandin[1], Séverine Kirchner[3] and André Cicolella[1]

[1]National Institute of Environment and Risks, France (INERIS)
[2]University of Technology of Compiègne, France (UTC)
[3]Observatory of Indoor Air Quality, France (OQAI)

[*]*Corresponding email: massonje@utc.fr*

SUMMARY

This study aims to identify the most relevant variables, among outdoor measures, building characteristics and socioeconomic situation, for predicting indoor air chemical pollution in dwellings. To achieve this, we propose a two-step plan: first, group the dwellings into classes according to the indoor measured concentrations, then use regression tools to express a dwelling's class as a function of the aforementioned variables. In the first step, we use model-based clustering algorithms in a multivariate Gaussian mixture context; in the second step, we use binary decision trees in a discrimination context. This approach does not handle the pollutants individually, but considers them together as a multidimensional variable that must be summarized by a categorical variable (the dwelling's class).

INTRODUCTION
The French Observatory of Indoor Air Quality conducted a national monitoring survey in 567 French dwellings between 2003 and 2005 (Kirchner et al., 2007). More than thirty parameters (chemical, biological, and physical) were measured during one week, indoors and outdoors. Simultaneously, detailed information about the dwellings' characteristics as well as their occupants' situations and activities was collected. This study aims to describe the relationship between indoor air quality in terms of chemical concentrations and variables accessible on geographic units over France (with census data, land use plans…). The latter include information on the dwelling (house or flat, building age, surface area…), its occupants (number, ages and sexes, professional occupations…), and its equipment (type of heating system, separate bathroom, communicating garage, number of vehicles…).

METHODS
In a first step, we use mixture models and the associated EM and CEM algorithms to divide the group of dwellings (called population thereafter) into groups representing different types of indoor air pollution. We restrain to the 20 variables corresponding to indoor chemical concentrations (listed in Table 1). This procedure provides a new qualitative variable, that associates each dwelling with its class according to the mixture model. The core assumption is that the dwellings belonging to the same class form an independent and identically distributed sample, whose probability distribution has a given form. EM is an optimization algorithm suited to find maximum likelihood estimators (Dempster et al., 1977); CEM is one of its variants, adapted to the problem of statistical classification given the number of classes (Celeux and Govaert, 1992). Estimating mixture model parameters (proportion of each class,

descriptors of the conditional distributions) this way permits the estimation of the unobserved class variable (McLachlan and Basford, 1988). The chosen models are the family of Gaussian parsimonious clustering models (Celeux and Govaert, 1995), and the "best" model (number of classes, form of the covariance matrices) is selected with the Bayesian Information Criterion (Schwarz, 1978). This way, we summarize the twenty indoor chemical concentrations by one categorical variable.

Table 1. List of the 20 measured chemicals (one measure indoors and one outdoors).

Code	Substance	Code	Substance
ald21	formaldehyde	cov47	ethylbenzene
ald22	acetaldehyde	cov48	m+p-xylene
ald23	acrolein	cov49	styrene
ald24	hexaldehyde	cov50	o-xylene
cov41	benzene	cov51	2-butoxy-ethanol
cov42	1-methoxy-2-propanol	cov52	124-trimethylbenzene
cov43	trichloroethylene	cov53	1,4-dichlorobenzene
cov44	toluene	cov54	n-decane
cov45	tetrachloroethylene	cov55	2-butoxy-ethyl-acetate
cov46	1-methoxy-2-propyl-acetate	cov56	n-undecane

In a second step, we use the classification and regression trees (CART) methodology to identify which of the explanatory variables most influence the class (Breiman et al., 1984). It consists in splitting the population into two subpopulations, then splitting each subpopulation into two parts, and so on. Each splitting is based on one explanatory variable, and chosen so that the subpopulations most differ according to the values of the explained variable (here, the class obtained in the first step). This method provides a binary decision tree, that can be used to predict the class of a dwelling from the knowledge of the explanatory variables. This way, we build an interpretation of the classes as profiles in terms of those variables.

Both steps are performed in the free statistical environment *R*, using respectively the packages *mclust* (Fraley and Raftery, 2007) and *tree* (by B. Ripley).

RESULTS
In the first step, 534 out of the 567 dwellings are classified, because of the existence of missing values: 33 incomplete individuals are removed before the analysis. The best model found by the *Mclust* function is "VEI with 12 components", which means "Varying volume, Equal shapes and Identity direction", and twelve classes. It is a diagonal model, where the covariance matrices are of the form $\lambda_k B$; λ_k is a scalar proper to each class k and B is a diagonal matrix. According to this model, the variables are independent conditionally to the classes, and the covariance matrices are all proportional. This clustering procedure's results are of two natures:
− Estimators of the model's parameters: 12 scale values (one for each class), 20 shape values (one for each variable), and 240 location values (the classes' means: one for each variable per class). Tables 2 and 3 summarize the scale and shape parameters, but for the sake of readability we do not reproduce here the location ones. Note that the model's theoretical variance of variable p for class k is equal to $\lambda_k b_p$.
− A classification corresponding to these parameters: each individual (dwelling) is assigned to the class to which it belongs with highest probability. These probabilities are calculated using the above-mentioned parameters. Table 4 summarizes the number n_k of dwellings assigned to each class k.

Table 2. Summary of step 1 results: scale parameters.

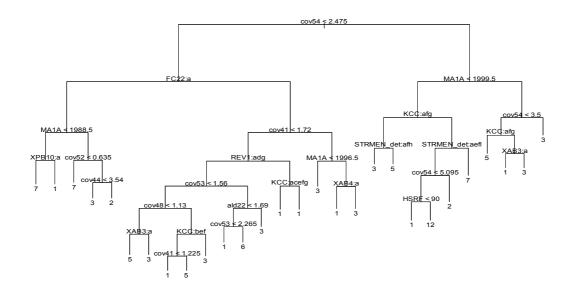| Class k	1	2	3	4	5	6
Scale λ_k	1.13	49.33	4.63	8577.12	36.59	1.25
Class k	7	8	9	10	11	12
Scale λ_k	9.57	782506.90	5.14	14.23	11.91	1549.43

Table 3. Summary of step 1 results: shape parameters (coefficients of *B*).

Variable p	ald21	ald22	ald23	ald24	cov41	cov42	cov43
Shape b_p	29.04	9.85	0.12	19.08	0.35	5.27	0.73
Variable p	cov44	cov45	cov46	cov47	cov48	cov49	cov50
Shape b_p	17.85	2.21	0.08	0.29	2.04	0.06	0.30
Variable p	cov51	cov52	cov53	cov54	cov55	cov56	
Shape b_p	0.80	1.47	65.80	12.51	2.5e-6	13.39	

Table 4. Summary of step 1 results: number of dwellings assigned to each of the 12 classes.

Class k	1	2	3	4	5	6	7	8	9	10	11	12
n_k	140	23	167	7	79	10	37	3	15	9	11	33

The binary decision tree produced at step 2 is shown in Figure 1. Each ramification of the tree is associated to a binary test on one of the explanatory variables. The root (at the top) corresponds to the population of the 534 studied dwellings, and each ramification splits the incoming population into a left-subpopulation (positive test) and a right-subpopulation (negative test). When it is no longer possible to split a population (if its size is too small, or if all its elements have the same class), the branch stops at a "leaf", which is then associated to the most frequent class in the corresponding population. For example, the leftmost leaf is associated with the individuals verifying:

− the outdoor measure of n-decane is lower than 2.475 µg.m^{-3},
− there is a communicating garage (*FC22* is a binary variable: yes[a] or no[b]),
− the reference occupant arrived in the dwelling in 1988 or before,
− there is a building site at a distance lower than 500m (*XPB10* is also a binary variable).

Figure 1. Discrimination tree obtained after step 2.

DISCUSSION

We presented here the two-step structure of our method, and some preliminary results. They show that our approach is practicable, but will need refinements because of several limitations. Notably, the classes 4, 8 and 12 obtained in step 1 are not useful for prediction because they are associated with very high values of λ_k (so the associated theoretical variances are very high too). The corresponding dwellings have to be studied apart to improve the classification. Indeed, all results reported here are subject to be either validated or changed, and they should be viewed as an illustration. However, we believe that the flexibility of the tools we use will permit us to overcome these limitations.

CONCLUSIONS

We have built a decision procedure to roughly assess the indoor air pollution in dwellings based on variables that are spatially available. This will enable us to produce maps of indoor air quality categories for French dwellings, where each category is associated with specific values for the parameters of the 20-dimensional Gaussian distribution followed by the indoor chemical concentrations. Many details must still be improved before drawing any definitive conclusions, but the method itself seems adapted to our research question.

REFERENCES

Breiman L., Freidman J., Olshen R., and Stone C. 1984. *CART: Classification and regression trees*. Wadsworth, Statistics/Probability series. Wadsworth.

Celeux G. and Govaert G. 1992. A classification EM algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3), 315–332.

Celeux G. and Govaert G. 1995. Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5), 781–793.

Dempster A., Laird N., and Rubin D. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1), 1–38.

Fraley C. and Raftery A. 2007. *MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering*. Technical Report No. 504, Department of Statistics, University of Washington.

Kirchner S., Arènes J.-F., Cochet C., Derbez M., Duboudin C., Elias P., Grégoire A., Jédor B., Lucas J.-P., Pasquier N., Pigneret M., and Ramalho O. 2007. État de la qualité de l'air dans les logements français. *Environnement, Risques et Santé*, 6(4), 259–269.

McLachlan G. J. and Basford K. E. 1988. *Mixture Models, Inference and Applications to Clustering*. Marcel Dekker.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.r-project.org/

Schwarz G. 1978. Estimating the dimension of a model. *The Annals of Statistics*, 6(2), 461–464.