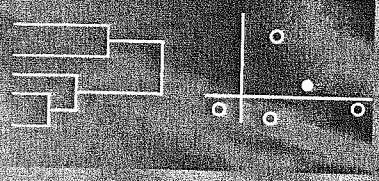
STUDIES IN CLASSIFICATION, DATA ANALYSIS, AND KNOWLEDGE ORGANIZATION

Data Analysis and Classification

Francesco Palumbo
Carlo Natale Lauro
Michael J. Greenacre
Editors



Springer

Table 1 Nitrogen dioxide data: value of rmspe(d) for the considered algorithms

d	LC-KF	KF	std
1	0.0920	0.1027	0.6508
2	0.1969	0.2102	0.6595
3	0.2950	0.3070	0.6659
4	0.3860	0.4052	0.6708

in the dynamic model strongly depends on the value of d. Second, the LC algorithm seems to perform better than the standard algorithm, for all the values of d we have considered.

6 Conclusions

In this work we show how bias effects of the covariate errors in the estimation process by Kalman Filter of a dynamic linear regression model can be understood in terms of the Löwner partial ordering. The application of the proposed LC-KF algorithm to a problem of Nitrogen Dioxide air concentration forecasting shows how the proposed method may mitigate the negative effects of the covariate errors by improving the predictive performances of a given model.

References

Carroll, R., Ruppert, D., Stefanski, L. A., & Crainiceanu, L. M. (2006). *Measurement error in nonlinear models* (2nd edition). Boca Raton: Chapman and Hall/CRC.

Chowdhury, S., & Sharma, A. (2007). Mitigating parmeter bias in hydrological modelling due to uncertainty in covariates. *Journal of Hydrology, 340*, 197–204.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association, 89*(428), 1314–1328.

Fuller, W. A. (1986). *Measurement error models*. New York: Wiley.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*, 35–45.

Mantovan, P., & Pastore, A. (2004). Flexible dynamic regression models for real-time forecasting of air pollutant concentration. In: H. H. Bock, M. Chiodi, & A. Mineo (Eds.), *Advances in multivariate data analysis* (pp. 265–276). Berlin: Springer.

Mantovan, P., Pastore, A., & Tonellato, S. (1999). A comparison between parallel algorithms for system parameter estimation in dynamic linear models. *Applied Stochastic Models in Business and Industry, 15*, 369–378.

Marshall, A. W., & Olkin, I. (1979). *Inequalities – Theory of majorization and its applications*. New York: Academic.

West, M., & Harrison, J. (1997). *Bayesian forecasting and dynamic models* (2nd edition). New York: Springer.

Local Multilevel Modeling for Comparisons of Institutional Performance

Simona C. Minotti and Giorgio Vittadini

Abstract We propose a general methodology for evaluating the quality of public sector activities such as education, health and social services. The traditional instrument used in comparisons of institutional performance is Multilevel Modeling (Goldstein, H., Multilevel statistical models, Arnold, London, 1995). However, rankings based on confidence intervals of the organization-level random effects often prevent to discriminate between institutions, because uncertainty intervals may be large and overlapped. This means that, in some situations, a single global model is not sufficient to explain all the variability, and methods able to capture local behaviour are necessary. The proposal, which is entitled Local Multilevel Modeling, consists of a two-step approach which combines Cluster-Weighted Modeling (Gershenfeld, N., The nature of mathematical modeling, Cambridge University Press, Cambridge, 1999) with traditional Multilevel Modeling. An example regarding the evaluation of the "relative effectiveness" of healthcare institutions in Lombardy region is discussed.

1 Introduction

In the 1990s, numerous authors proposed the use of Multilevel Models (Goldstein 1995) in institutional comparisons (see Bryk and Raudenbush 2002, Chap. 5), with rankings based on confidence intervals of the random effects associated with organizations. However, "an overinterpretation of a set of rankings where there are large uncertainty intervals can lead both to unfairness and inefficiency and unwarranted conclusions about changes in ranks" (Goldstein and Spiegelhalter 1996). This is the case of regional or national studies, where confidence intervals may be large and overlapped due to the heterogeneity of individuals within organizations and, more specifically, to non-homogeneity and non-linearity of individual relationships.

S.C. Minotti (✉)
Dipartimento di Statistica, Università degli Studi di Milano-Bicocca, Via Bicocca degli Arcimboldi 8, 20126 Milano, Italy
e-mail: simona.minotti@unimib.it

We argue that a single global model is not sufficient to interpret such a complexity, and methods able to capture local behaviour are necessary. We will show how the combination of Cluster-Weighted Modeling and traditional Multilevel Modeling represents a very flexible methodology for comparisons of institutional performance. The proposal, which is entitled Local Multilevel Modeling, provides local rankings that allow the policy makers to identify institutions which are above or below the average for specific groups of users.

Typical methods able to detect and model unknown patterns at individual-level are described in Sect. 2. The proposal of Local Multilevel Modeling will be introduced in Sect. 3. Then, in Sect. 4, an example regarding the evaluation of the 'relative effectiveness' of healthcare institutions in Lombardy region is discussed. Finally, in Sect. 5 we provide conclusions and discuss further research.

2 Capturing Local Behaviour

2.1 Mixture Modeling

A good approach to capture local behaviour at individual-level is given by Mixture Modeling (McLachlan and Basford 1988), which is closely related to splitting up a data set by clustering and is an example of unsupervised learning.

Given a set of multivariate data points $\{\mathbf{x}_n\}_{n=1}^N$, where $\mathbf{x}$ is a vector of real-valued predictors, Mixture Modeling factors the density $p(\mathbf{x})$ over multivariate class-conditional densities:

$$p(\mathbf{x}) = \sum_{k=1}^{K} p(\mathbf{x}, c_k) = \sum_{k=1}^{K} p(\mathbf{x}|c_k)p(c_k), \tag{1}$$

where $p(\mathbf{x}|c_k)$, $(k = 1, \ldots, K)$, is the k-th component density and is assumed to be multivariate Gaussian, and the prior probability $p(c_k)$, $(k = 1, \ldots, K)$, is the k-th mixing parameter (i.e. the fraction of the data explained by cluster c_k).

The posterior probability $p(c_k|\mathbf{x})$, $(k = 1, \ldots, K)$, which is given by

$$p(c_k|\mathbf{x}) = \frac{p(\mathbf{x}, c_k)}{p(\mathbf{x})} = \frac{p(\mathbf{x}|c_k)p(c_k)}{\sum_{k=1}^{K} p(\mathbf{x}|c_k)p(c_k)}, \tag{2}$$

indicates the fraction of a point to be associated with the k-th Gaussian; "as a Gaussian begins to have a high probability at a point, that point effectively disappears from the other Gaussians" (Gershenfeld 1999).

However, Mixture Modeling is not completely adequate to the problem under study, because it does not take into account the functional dependence between individual-level characteristics $\mathbf{x}$ and the outcome variable y; the n-th individual, in fact, contributes to each cluster k, $(k = 1, \ldots, K)$, with a posterior probability $p(c_k|\mathbf{x}_n)$ which depends on individual characteristics $\mathbf{x}_n$ only.

2.2 Cluster-Weighted Modeling

A better alternative is given by Cluster-Weighted Modeling (Gershenfeld 1999), which is a framework for supervised learning that combines a density estimation of the input data with a functional relationship to the output data. This leads to a number of local clusters, each containing its own model for describing the observed data.

Cluster-Weighted Modeling essentially tries to estimate the joint density $p(y, \mathbf{x})$ by means of a Gaussian Mixture Model. "But where conventional Gaussian Mixture Modeling only estimates the quantity $p(\mathbf{x})$, Cluster-Weighted Modeling includes an additional output term $p(y|\mathbf{x}, c_k)$ to capture the functional dependence of the output values y_n on the input vectors $\mathbf{x}_n$ as part of the density estimation" (Engster and Parlitz 2006).

Given a set of multivariate data points $\{y_n, \mathbf{x}_n\}_{n=1}^N$, where y is a scalar (the generalization to vector $\mathbf{y}$ is straightforward) and $\mathbf{x}$ is a vector of real-valued predictors, Cluster-Weighted Modeling factors the joint density $p(y, \mathbf{x})$ over multivariate class-conditional densities:

$$p(y, \mathbf{x}) = \sum_{k=1}^{K} p(y, \mathbf{x}, c_k) = \sum_{k=1}^{K} p(y, \mathbf{x}|c_k)p(c_k) = \sum_{k=1}^{K} p(y|\mathbf{x}, c_k)p(\mathbf{x}|c_k)p(c_k), \tag{3}$$

where $p(y|\mathbf{x}, c_k)$, $(k = 1, \ldots, K)$, indicates a dependence in the output space and is assumed to be multivariate Gaussian (with mean given by a function which reflects prior belief on the local relationship between $\mathbf{x}$ and y, for example locally linear (Gershenfeld 1999)); $p(\mathbf{x}|c_k)$, $(k = 1, \ldots, K)$, indicates a domain of influence in the input space and is also assumed to be multivariate Gaussian; $p(c_k)$, $(k = 1, \ldots, K)$, is the k-th prior probability.

If there are unordered, discrete-valued variables $\mathbf{x}_d$ in addition to real-valued predictors $\mathbf{x}_r$, the prior probability $p(c_k)$ in (3) is conditioned on $\mathbf{x}_d$ and hence is replaced by $p(c_k|\mathbf{x}_d)$. Moreover, if y indicates a target variable (in classification problems), the output term in (3) is simply a histogram of the probability to see each state for each cluster and we refer to Cluster-Weighted Classification (Gershenfeld 1999).

The posterior probability $p(c_k|\mathbf{x})$, $(k = 1, \ldots, K)$, is given by

$$p(c_k|y, \mathbf{x}) = \frac{p(y, \mathbf{x}, c_k)}{p(y, \mathbf{x})} = \frac{p(y|\mathbf{x}, c_k)p(\mathbf{x}|c_k)p(c_k)}{\sum_{k=1}^{K} p(y|\mathbf{x}, c_k)p(\mathbf{x}|c_k)p(c_k)}. \tag{4}$$

Analogously to Mixture Modeling, the parameters are estimated using an expectation maximization (EM) algorithm, leading to a local optimum in parameter space.

Cluster-Weighted Modeling is adequate to the problem under study, because it does take into account the functional dependence between user-level characteristics $\mathbf{x}$ and the outcome y; the n-th individual, in fact, contributes to each cluster k, $(k = 1, \ldots, K)$, with a posterior probability $p(c_k|y_n, \mathbf{x}_n)$ which depends on both y_n and $\mathbf{x}_n$.

3 The Proposal: Local Multilevel Modeling

We propose a general methodology for local comparisons of institutional performance. The proposal consists of a two-step approach which combines Cluster-Weighted Modeling with traditional Multilevel Modeling.

During the first step, we model the heterogeneity of individuals by means of Cluster-Weighted Modeling, by taking as response variable a continuous outcome y and as predictors the individual-level characteristics $\mathbf{x}$:

$$p(y, \mathbf{x}) = \sum_{k=1}^{K} p(y, \mathbf{x}, c_k) = \sum_{k=1}^{K} p(y, \mathbf{x}|c_k)p(c_k) = \sum_{k=1}^{K} p(y|\mathbf{x}, c_k)p(\mathbf{x}|c_k)p(c_k), \tag{5}$$

where $p(y|\mathbf{x}, c_k)$, $p(\mathbf{x}|c_k)$ and $p(c_k)$ are defined in (3).

If we assign each individual to the cluster with the highest posterior probability, groups of users characterized by the same relationship between individual-level characteristics and the outcome are identified.

In the second step, we apply a Multilevel Model for each cluster k, $(k = 1, \ldots, K)$, by taking as response variable the outcome y and as predictors the individual-level and organization-level characteristics, $\mathbf{x}$ and $\mathbf{z}$ respectively:

$$y_{ij} = \alpha_j + \sum_{g=1}^{G} \beta_{gj} x_{gij} + e_{ij}, \qquad \alpha_j = \sum_{h=1}^{H} \delta_{hj} z_{hj} + u_j, \tag{6}$$

where y_{ij} is an outcome regarding the i-th individual $(i = 1, \ldots, n_j; N = n_1 + \cdots + n_j + \cdots + n_Q)$ belonging to the j-th organization $(j = 1, \ldots, Q)$, α_j is a random coefficient associated with the j-th organization, β_{gj} is a fixed coefficient associated with individual-specific covariate x_{gij}, $(g = 1, \ldots, G)$, e_{ij} is a random disturbance associated with the i-th individual belonging to the j-th organization; δ_{hj} is a fixed coefficient associated with organization-specific covariate z_{hj}, $(h = 1, \ldots, H)$, u_j is a random effect associated with the j-th organization, adjusted for individual-level and organization-level characteristics.

The plot of confidence intervals for ordered second-level random effects provides, for each cluster k, a ranking of organizations. The advantage is that Local Multilevel Modeling gives to the policy makers an instrument which is able to locally highlight differences among institutions, i.e. for specific groups of users.

4 An Example

We briefly discuss an example regarding the evaluation of the "relative effectiveness" of healthcare institutions in Lombardy region. The data involved in the study refer to a sample of 22,877 patients hospitalised in 168 hospitals in 2006 and are provided by the Regional Agency for Health Care. The variables utilised in the study are described in Table 1, where y is the outcome, x_1–x_7 denote the patient-level characteristics (1), while z_1–z_3 the hospital-level characteristics (2).

In Table 2 the odds ratios for the significative variables in the reference Logistic Multilevel Model are reported.

The results indicate that the odds of death considerably increases for hospitals with large levels of high medical case-mix and low medical case-mix, and increases also for patients with emergency and tumour diagnosis. Note that the null effect of the age variable is due to the particular nature of the sample, which is merely composed by old patients (Mode = 71 years, Median = 60 years, Mean = 55.39 years). For sake of brevity we do not include the figure regarding hospital ranking, which shows that confidence intervals of the second-level random effects are large and overlapped.

Table 1 Variables in the effectiveness study

Label	Variable	Type
y	Total mortality indicator	Dichotomous
x_1	Age (1)	Discrete; number of years
x_2	Comorbidity (1)	Discrete; six levels
x_3	Length of stay (1)	Discrete; number of days
x_4	Relative level of case severity (1)	Continuous
x_5	Cardiovascular diagnosis (1)	Dichotomous
x_6	Emergency diagnosis (1)	Dichotomous
x_7	Tumour diagnosis (1)	Dichotomous
z_1	Low surgical case-mix (2)	Continuous
z_2	High medical case-mix (2)	Continuous
z_3	Low medical case-mix (2)	Continuous

Table 2 Reference model. Odds ratios for the significative variables (p-value < 0.05)

Variable	Beta	OR	Std. error
Age (1)	0.060	1.063	0.012
Emergency diagnosis (1)	1.340	3.822	0.416
Tumour diagnosis (1)	1.182	3.262	0.384
High medical case-mix (2)	2.199	9.019	0.370
Low medical case-mix (2)	1.728	5.630	0.858

Table 3 Cluster 1. Emergency and tumour diagnosis frequency tables

	y = 0, 1		y = 0		y = 1	
	Frequency	%	Frequency	%	Frequency	%
No emergency	20,379	98.62	20,322	99.45	57	24.89
Emergency	285	1.38	113	0.55	172	75.11
No tumour	19,191	92.87	18,997	92.96	194	84.72
Tumour	1,473	7.13	1,438	7.04	35	15.28
Total	20,664	100.00	20,435	100.00	229	100.00

Table 4 Cluster 2. Emergency and tumour diagnosis frequency tables

	y = 0, 1		y = 0		y = 1	
	Frequency	%	Frequency	%	Frequency	%
No emergency	645	87.52	301	90.39	344	85.15
Emergency	92	12.48	32	9.61	60	14.85
No tumour	373	50.61	58	17.42	315	77.97
Tumour	364	49.39	275	82.58	89	22.03
Total	737	100.00	333	100.00	404	100.00

Table 5 Cluster 3. Emergency and tumour diagnosis frequency tables

	y = 0, 1		y = 0		y = 1	
	Frequency	%	Frequency	%	Frequency	%
No emergency	493	33.40	73	6.97	420	98.13
Emergency	983	66.60	975	93.03	8	1.87
No tumour	1,278	86.59	983	93.80	295	68.93
Tumour	198	13.41	65	6.20	133	31.07
Total	1,476	100.00	1,048	100.00	428	100.00

The main purpose of the example is to show how the results of Cluster-Weighted Modeling allow us to identify subpopulations of patients. In particular, three is the optimal number of clusters derived from AIC and BIC criteria. If each patient is assigned to the cluster with the highest posterior probability, we have 20,664 patients in cluster 1, 737 in cluster 2 and 1,476 in cluster 3. Some of the main characteristics of the three clusters are briefly described in the following (see Tables 3, 4 and 5).

Cluster 1 is characterized by a mortality rate equal to 1.11% (against the 4.64% in the entire sample); the 75.11% of the dead patients had an emergency diagnosis (against the 22.62% in the entire sample). Instead Cluster 2 is characterized by an high percentage of patients with tumour diagnosis (49.39% vs. 8.90% in the entire sample) and in particular which are not dead (the 82.58% of the living patients, against the 8.15% in the entire sample); moreover, the living patients had high levels of comorbidity. Here the mortality rate is 54.82%, where the 77.97% of the dead patients had no tumour diagnosis (near to the percentage of 75.78% in the entire sample). By the end, cluster 3 is characterized by patients with emergency diagnosis (66.60% vs. 5.94% in the entire sample) and in particular which are not dead (93.03% of the living patients, against 5.13% in the entire sample). Here the

Table 6 Local models. Odds ratios for the significative variables (p-value < 0.05)

Variable	Beta	OR	Std. error
Age (1)	0.065	1.067	0.038
Cardiovascular diagnosis (1)	−0.240	0.786	0.116
High medical case-mix (2)	2.239	9.389	0.171
Age (1)	0.066	1.068	0.040
Comorbidity (1)	0.177	1.195	0.074
Emergency diagnosis (1)	1.528	4.609	0.696
High medical case-mix (2)	2.607	13.558	0.861
Age (1)	0.055	1.057	0.033
Comorbidity (1)	0.164	0.089	1.178
Emergency diagnosis (1)	1.540	4.668	0.582
Tumour diagnosis (1)	1.205	3.335	0.313
High medical case-mix (2)	2.074	7.960	0.428
Low medical case-mix (2)	1.755	5.784	0.416

mortality rate is 29%, where the 98.13% of the dead patients had no emergency diagnosis (against the 77.38% in the entire sample).

In Table 6 the odds ratios for the significative variables in the three Local Logistic Multilevel Models are reported.

For sake of brevity we do not include the ranking of hospitals for each group of patients, where confidence intervals of the hospital-level random effects are less large and overlapped than in the entire sample.

The example shows that there are, in this case, three different populations of patients which require three different models. The choice of a single global model would have caused a large loss of information. For example, hospital-level variable z_2, high medical case-mix, is an important variable for both the reference and the local models. However, the much higher effect in the second model seems to indicate that patients assigned to cluster 2 are characterized by a higher level of complexity, as confirmed by previous analysis of frequency tables. Analogous considerations can be made regarding the other variables.

5 Conclusions and Further Research

We have proposed a general methodology for the evaluation of public sector activities, which is entitled Local Multilevel Modeling. The proposal applies to regional or national studies, where a single global model is not sufficient to describe situations typically characterized by heterogeneity. The idea underlying our proposal is that non-homogeneity and non-linearity may appear in the individual-level relationships. The two-step approach proposed is able, firstly, to capture and model individual-level relationships whatever they are and, secondly, to locally highlight differences among organizations, i.e. for specific groups of users.

Of course, precautions which are valid for the use of traditional Multilevel Modeling in institutional comparisons hold true also for Local Multilevel Modeling. First, "we should exert caution when applying statistical models to make comparisons between institutions, treating results as suggestive rather than definitive" (Goldstein and Spiegelhalter 1996). Secondly, "measurement of outcomes for research purposes is useful to help organisations to detect trends and spot extreme outliers, but league tables of outcomes are not a valid instrument for day-to-day performance management by external agencies" (Lilford et al. 2004).

Further research will regard the extension of Cluster-Weighted Modeling to multilevel data structures, as proposed in Galimberti and Soffritti (2007) for Mixture Models, in order to allow some of the parameters of the conditional densities to differ across second-level units (schools, hospitals, etc.). This would reinforce the first step of the proposal, by taking into account also eventual non-homogenous and non-linear second-level relationships.

Acknowledgements The authors would like to express their thanks to Maurizio Sanarico for his valuable advice.

References

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models. Applications and data analysis methods*. Newbury Park, CA: Sage.

Engster, D., & Parlitz, U. (2006). Local and cluster weighted modeling for time series prediction. In B. Schelter, M. Winterhalder, & J. Timmer (Eds.), *Handbook of time series analysis. Recent theoretical developments and applications* (pp. 39–65). Weinheim: Wiley.

Galimberti, G., & Soffritti, G. (2007). Multiple cluster structures and mixture models: Recent developments for multilevel data. In *Book of short papers CLADAG 2007 "Meeting of the Classification and Data Analysis Group of the Italian Statistical Society"* (pp. 203–206), September 12–14, Università degli Studi di Macerata. EUM, Macerata.

Gershenfeld, N. (1999). *The nature of mathematical modeling*. Cambridge: Cambridge University Press.

Goldstein, H. (1995). *Multilevel statistical models*. London: Arnold.

Goldstein, H., & Spiegelhalter, D. J. (1996). League tables and their limitations: Statistical issues in comparisons of institutional performance. *JRSS A, 159*(3), 385–443.

Lilford, R., Mohammed, M. A., Spiegelhalter, D. J., & Thomson, R. (2004). Use and misuse of process and outcome data in managing performance of acute medical care: Avoiding institutional stigma. *The Lancet 363*, 1147–1154.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering*. New York: Marcel Dekker.

Modelling Network Data: An Introduction to Exponential Random Graph Models

Susanna Zaccarin and Giulia Rivellini

Abstract A brief introduction to statistical models for complete network data is presented. An example is provided by the collaboration network of Italian scholars on Population Studies.

1 Introduction

Social network data can be viewed as a social relational system characterized by a set of actors – maybe with their attribute variables – and their social ties (Wasserman and Faust 1994). Two main types of social network data are distinguished: ego-centered or personal network, and complete or one-mode network. Ego-centered network data are usually collected from a sample of actors (egos) reporting on the ties with and between other people (alters). Complete network data, on the other hand, concern a well-defined group of actors who report on their ties with all other actors in the group.

This contribution provides an introductory outline for statistical modelling of relational data, summarizing in particular the current methodological developments of the exponential random graph models (p^*) for complete social networks (Sect. 2).

The statistical models applied in social network analysis are typically non-standard because the common assumption of independent observations does not hold: the multiple ties to and from the same actor are related (Rivellini and Zaccarin 2007). Moreover, the popular assumption of continuous normally distributed variables does not hold when tie variables are binary, nominal, ordinal, or count variables.

As an example of fitting such a model, data from the collaboration network of Italian scholars in Population Studies for the year 2001 will be used (Sects. 3–5). The description of network and node properties by the means of the usual network measures will also be given.

S. Zaccarin (✉)
Università di Trieste, Piazzale Europa 1, 34127 Trieste, Italy
e-mail: susanna.zaccarin@econ.units.it