First CLADAG Data Mining Prize: Data Mining for Longitudinal Data with Different Marketing Campaigns

Mouna Akacha, Thaís C. O. Fonseca and Silvia Liverani

Abstract The CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society recently organised a competition, the 'Young Researcher Data Mining Prize' sponsored by the SAS Institute. This paper was the winning entry and in it we detail our approach to the problem proposed and our results. The main methods used are linear regression, mixture models, Bayesian autoregressive and Bayesian dynamic models.

1 Introduction

Recently the CLAssification and Data Analysis Group (CLADAG) of the Italian Statistical Society organised a competition on a data mining problem. Given the sales of nine products over seven time periods, five structural variables and a marketing campaign treatment for 4517 sales points, the competitors are asked to

1. evaluate the marketing campaign impact on the economic return in the first time period,
2. and forecast the economic return for the seventh time period.

The first question may be seen as a problem studied in regression analysis, whilst the second problem is widely studied in time series forecasting. However, the presence of several covariates with non-linear and co-dependent features requires both questions to be addressed with ad hoc methods. The main statistical method we use to address the first question is a mixture of regression models, while we fit autoregressive and dynamic models for the forecasting problem.

Department of Statistics, University of Warwick, UK. Mouna Akacha e-mail: m.akacha@warwick.ac.uk · Thais C. O. Fonseca e-mail: t.c.o.fonseca@warwick.ac.uk · Silvia Liverani e-mail: s.liverani@warwick.ac.uk. The three authors contributed equally to this project and should be considered co-first authors.

This paper is organised as follows. In Section 2 we describe the data and perform exploratory tests and analysis. In Section 3 we detail the mixture model and our results to answer the first question asked by the organisers. In Section 4 we introduce the autoregressive and dynamic models for prediction and present our results for the second question.

2 The Data

The data provided in this competition was collected by sales points over seven time periods. The outcome variable y_{it} is an unknown function of the income of the sales point i during time period t, with $t = 1,\dots,7$ and $i = 1,\dots,n$ where $n = 4517$. We define $y_t = (y_{1t},\dots,y_{nt})'$. Five structural variables are available for each sales point, $x_{1i},\dots,x_{5i}$. They are time invariant and they have 2, 3, 4, 2 and 3 levels respectively. For each time period t the sales for nine products are available. The nine product sales are defined as $s_{jt}^{(i)}$ where $j = 1,\dots,9$ is the product index, t is the time period and i is the unit. For simplicity, we refer to the product j at time t for all units as the vector $s_{jt} = (s_{jt}^{(1)},\dots,s_{jt}^{(n)})$. Finally, only some of the sales point have received a certain marketing campaign during the first time period. The marketing campaign indicator $z_i = 0$ if the sales point i is in the control group, and one otherwise. We will use the terms marketing campaign and treatment interchangeably.

From now on, to simplify the notation, we will use x_k, for $k = 1,\dots,5$ to represent the 4517-dimensional vector of the values of the kth structural variable. We will use an analogous notation for z. Like for most data mining tasks, we now have to explore the data before fitting any model.

Fig. 1(a) shows y_1 against s_{11}. Note the shrinkage of s_{11} towards the left due to the large range of its tails. Analogous plots for variables s_{j1}, for $j = 2,\dots,9$, omitted here, show similar patterns. In order to improve the spread of the data, we use the log transformation on the product sales s_{jt}. This is common practice in the context of sales and prices. For example, Fig. 1(b)-(c) show the result of the log transformation of s_{11} against y_1. Note the strong pattern formed by the marketing treatment z and structural variable x_3 as shown in these plots. Note that there are clusters of sales points that have similar outcomes, similar sales of product 1 and the same value of the structural variable x_3 too. Similar plots, omitted here, can be obtained by conditioning on the levels of other structural variables and marketing campaign, so the data seems highly structured.

However, an issue arises when we apply the log transformation: there are several products with zero sales observed. For example, during the first time period there are 609 sales points that have zero sales for at least one product. Table 1 shows the number of units with product sales equal to zero for $t = 1$. For plotting we assigned $\log s_{jt}^{(i)} = -1$ when $s_{jt}^{(i)} = 0$. Due to the high frequency of such cases we will propose a model that accommodates this feature in Section 3. Moreover, note the highly skewed distribution of the product sales s_{it} for the sales points that do not sell all of the nine products in a time period, for example, for s_{51} in Fig. 2(a).

Table 1 Number of units with product sales equal to zero observed for $t = 1$.

	s_{11}	s_{21}	s_{31}	s_{41}	s_{51}	s_{61}	s_{71}	s_{81}	s_{91}	total
Counts	134	37	0	0	43	150	142	115	96	717

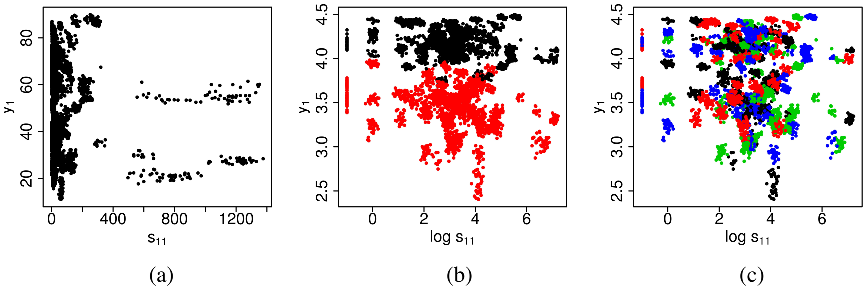


Fig. 1 Scatter plots for the outcome and the product sales. The units are coloured in different colours in plot (b) depending on the treatment z and in plot (c) depending on the levels the structural variable x_3.

The plots above revealed the presence of an association between the sales y_1, the treatment campaign z and the structural variables x's, but note also the association between the sales of two different products during the same time period, shown in Fig. 2(b) for the first time period. This relation is reasonably linear, even though clusters generated by the structural variables are still very clear.

The scenario changes dramatically when we move from the first time period, which seems strongly affected by the marketing campaign, to the other time periods. Fig. 1(c) and Fig. 2(c) represent y_t against $\log(s_{1t})$ for $t = 1$ and $t = 2$ respectively: the structure, clearly visible in Fig. 1(c), is not apparent anymore in Fig. 2(c). Analogously, the association between product sales s_{j1} shown in Fig. 2(b) for the first time period is retained by the product sales s_{j2}, as shown in Fig. 2(d), but the pattern defined by structural variable x_3 for the first time period in Fig. 2(b) is not present anymore for the second time period in Fig. 2(d).

A strong correlation is also apparent for the outcome at sequential time periods. See, for example, the plots of y_2 against y_1 and y_3 against y_2, as shown in Fig. 2(e)-(f), conditionally on the marketing campaign in Fig. 2(e) and on x_3 in Fig. 2(f). A similar behaviour is apparent for the product sales. See Fig. 2(g) for s_{13} against s_{12}. A different outlook on this strong autoregressive component over time, and the dependence of the product sales and outcome on structural variables, is also given by the plots of the time series for 50 sales of product s_{1t} against time t in Fig. 2(h) and for the outcome y_t of 50 sales points against time in Fig. 2(i).

The boxplots in Fig. 2(j)-(l) show the time dependence of the impact of the marketing campaign z and structural variables x_1 and x_4 respectively on the outcome y_t. The effect of z and x_1 on y show a strong time dependence while this is not apparent for the other structural variables x_2, …, x_5, as shown by Fig. 2(l) and other plots omitted here.

One of the main issues with the dataset is due to the design of the marketing campaign: for the sales points in the control group not all the possible configurations of the structural variables x_i have been observed. In particular, there are 61 (out of 144) combinations of the categories of the structural variables x for which we have no information for $z = 0$. This accounts for more than a third of the possible configurations and it will affect our results by restricting our ability to test for the effect of some factors on the impact of the marketing campaign. Moreover, it should be noted that there are also over 25 configurations of the structural variables that had only 3 or less observations. Table 2 shows, for example, five of the configurations with low counts.

Table 2 Examples of configurations of the structural variables and their counts for different values of z.

	$z = 0$	$z = 1$
$x_1 = 1, x_2 = 1, x_3 = 2, x_4 = 2, x_5 = 1$	0	29
$x_1 = 1, x_2 = 1, x_3 = 2, x_4 = 2, x_5 = 2$	0	50
$x_1 = 1, x_2 = 2$	0	834
$x_1 = 1, x_3 = 2$	0	607
$x_1 = 2, x_2 = 1, x_3 = 3, x_4 = 1, x_5 = 2$	9	1

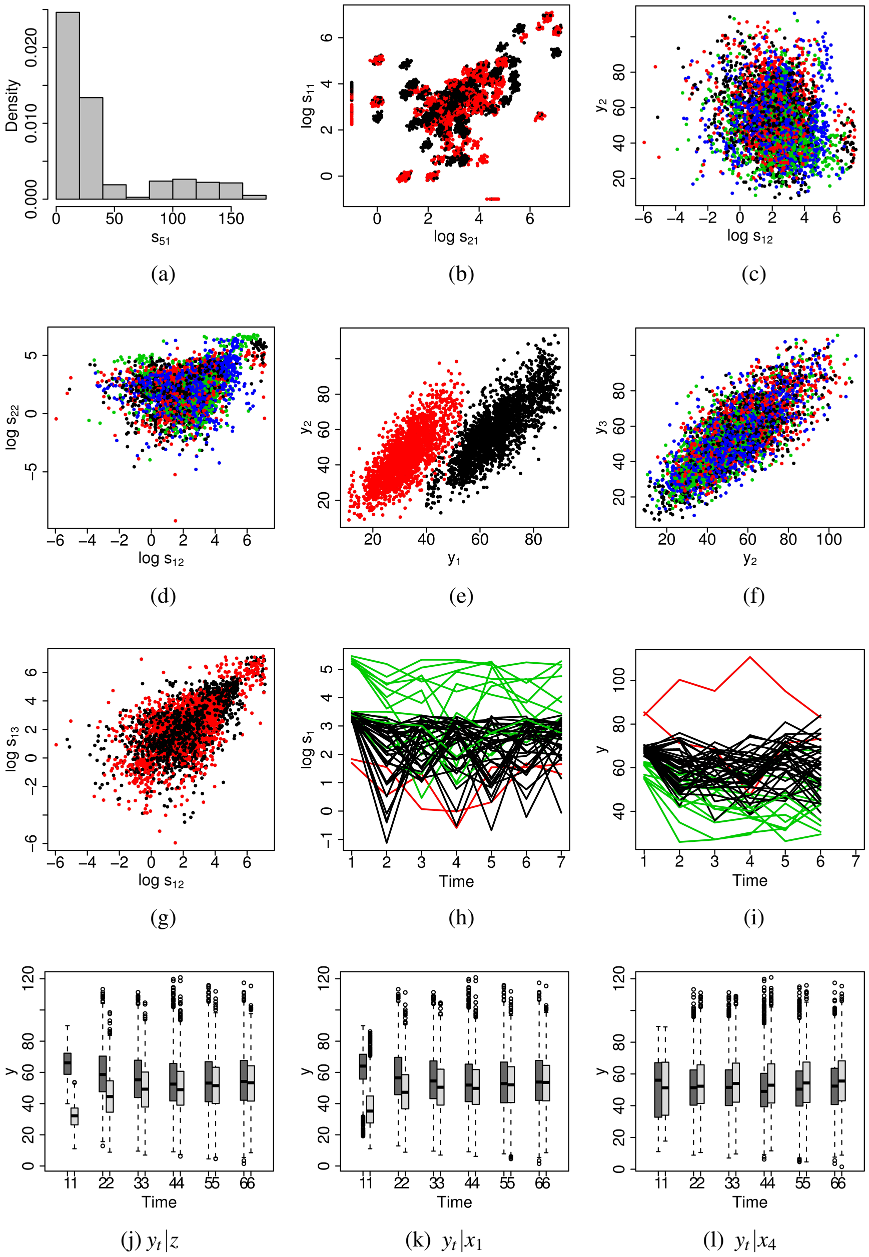
Another important feature of the data is the strong association between the structural variables. The null hypothesis of independence between pairs of structural variables is strongly rejected for certain pairs. The results of the χ^2 contingency test for the pairs (x_i, x_j), where

$$H_0 : x_i \perp x_j, \text{ for } i, j = 1, \ldots, 5 \text{ and } i \neq j,$$

are shown in Table 3. In order to reduce the complexity of the model, we choose the significance level 0.01. The result leads us to believe that a dependence structure between the structural variables exists and it is represented in Fig. 3.

Table 3 Approximated p-values for the χ^2 contingency test for H_0: $x_i \perp x_j$, for $i, j = 1, \ldots, 5$.

$\perp$	x_2	x_3	x_4	x_5
x_1	0.26	0.33	0.02	0.02
x_2		≤ 0.0001	≤ 0.0001	0.69
x_3			≤ 0.0001	≤ 0.0001
x_4				0.08



(a)

(b)

(c)

(d)

(e)

(f)

(g)

(h)

(i)

(j) $y_t|z$

(k) $y_t|x_1$

(l) $y_t|x_4$

Fig. 2 Histogram (a) shows the distribution of s_{51} for the sales points that have zero sales for at least one of the products. Plots (b)-(g) are scatter plots and time series plots for the outcome and the product sales. The colours depend on the levels of z, x_3, x_5, z, x_3, x_1, x_5 and x_5 respectively. The last time point in (i) has to be predicted. The boxplots in (j)-(l) show the evolution in time of the distribution of $y_t|z$, $y_t|x_1$ and $y_t|x_4$ respectively.

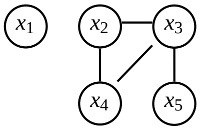


Fig. 3 Undirected graphical model to summarise the dependence structure obtained by the χ^2 contingency test shown in Table 3. The absence of an edge between two variables represents the lack of association between them.

Analogously, the product sales are also highly correlated. See Fig. 2(b) and (d). This does not allow us to include these covariates directly in the model matrix as we need it to be full rank. Thus we choose to use Principal Component Analysis (PCA), a dimensionality reduction method that transforms correlated continuous variables into a possibly smaller number of uncorrelated variables called principal components. In particular, here we implement the 'projection pursuit' method developed by [5] and based on robust estimation of covariance matrices. This method is especially useful for high dimensional data as ours. We apply PCA to the product sales s_{jt} and identify the principal components c_{kt} for $k = 1,\dots,9$.

The exploratory analysis of the data provided by the organisers uncovered the presence of a structure between the covariates and the outcome variable, and this structure provides us with the empirical motivation for the assumptions of the models proposed in Sections 3 and 4. However, it also uncovered issues that require careful consideration in the modeling stage, such as an unbalanced design and a strong association between some of the covariates.

3 The Impact of the Marketing Campaign on Outcome at Time Period 1

The first question asked by the organisers is to evaluate the impact of the marketing campaign z on the outcome y_1. As this question involves only the data for the first time period, in this Section we drop the subscript t and use the notation $s_1,\dots,s_9$ and y instead of $s_{11},\dots,s_{91}$ and y_1.

Regression models provide us with tools to identify and estimate the impact of treatment, that is,

$$y_i = \alpha_0 + \alpha_1 z_i + \text{error}_i, \tag{1}$$

where z_i is the factor indicating 1 for treatment and 0 for control for unit i. In particular, this model states that there is a change of magnitude α_1 in the mean outcome due to the marketing campaign. Fitting this linear model following the linear model proposed by [3] gave us the following estimates: $y_i = 32.0950 + 33.8743\, z_i$. The effect of treatment is highly significant with a p-value smaller than 0.0001 and adjusted $R^2 = 0.7742$. The coefficient of determination, R^2, is the proportion of variability in a data set that is accounted for by the statistical model. It provides a measure of how well future outcomes are likely to be predicted by the model. The adjusted R^2 is a modification of the coefficient of determination that adjusts for the number of explanatory terms in a model.

Although the residual analysis for this model seems reasonable, the availability of many covariates, as usual in a data mining problem such as this one, allows us to study the impact of the marketing campaign once the confounding effect of the other covariates has been removed.

The structural variables are potential covariates of interest, as shown in Section 2. Interaction terms will also need to be included in our model fit. However, the existence of empty cells in the design of the experiment on the marketing campaign (see Table 2) restricts the inclusion of all the possible interactions.

The product sales are potential covariates of interest as well. However, there is a high frequency of units with product sales equal to zero, as discussed in Section 2. For example, for the first time period, as shown in Table 1, there are 717 zero cells. This motivates our proposal of a mixture model, given by

$$y_i = \xi g_{1i} + (1-\xi) g_{2i} \qquad \text{with } \xi \in [0,1], \tag{2}$$

where

$$g_{1i} = \begin{cases} \mu_{1i} + \varepsilon_1 \text{ with } \varepsilon_1 \sim \mathscr{N}(0,\sigma_1^2), & \text{if } s_j^{(i)} > 0, \forall j = 1,\dots,9, \\ 0 & \text{otherwise,} \end{cases} \tag{3}$$

and

$$g_{2i} = \begin{cases} \mu_{2i} + \varepsilon_2 \text{ with } \varepsilon_2 \sim \mathscr{N}(0,\sigma_2^2), & \text{if } \exists\, s_j^{(i)} = 0, j = 1,\dots,9, \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

The MLE estimate of the proportion of products that are not sold during the first time period is given by $\hat{\xi} = 0.1348$ (with $SE(\hat{\xi}) = 0.0051$), a proportion of the data that cannot be ignored. Therefore, we propose different models for g_1 and g_2, to which we will refer as *group 1* and *group 2* respectively from now on. Further analysis, not included here, also confirmed a different variance between the two groups, justifying $\sigma_1^2 \neq \sigma_2^2$.

In the next subsections we perform stepwise model selection by AIC [8] to provide a measure of how well future outcomes are likely to be predicted by the model and to fit models to exclude non significant variables and interactions. .

3.1 Model for Group 2

The unbalanced design of the marketing campaign imposes restrictions on the interaction terms that can be included in the model for group 2. For instance, the interaction between x_1, x_2 and z cannot be included in the model as all the observed units with $x_1 = 1$ and $x_2 = 2$ were in the control group $z = 0$.

Note that the chi-square test performed in Section 2 brought to our attention the possibility of a presence of multicollinearity between several pairs of structural variables. It would be possible to check for collinearity by investigating the condition index [1], and then use e.g. correspondence analysis [2]. However, [9] argues that multicollinearity does not actually bias results, it just produces large standard errors in the related independent variables. With enough data, these errors will be reduced, and given the size of the dataset we therefore refrain from doing additional analyses. Moreover, an additional advantage of using raw data in the regression model is the easy interpretability of the parameter estimates.

Fitting models using a stepwise model selection by AIC to exclude non significant variables and interactions yields the final model which includes: z, x_1, ..., x_5, the interaction term (x_1, x_4), 7 indicator functions for the product sales s_j with $j = 1,2,5,6,7,8,9$ (the indicator is equal to 1 when products of that category have been sold in the first time period), $\log(s_3)$, $\log(s_4)$ and two indicator functions for $s_5 > 75$ and for $s_7 > 75$. Note that all the sales points sold a positive amount of products j for $j = 3,4$. Also, note that product sales s_5 and s_7 have a highly skewed distribution for the observations in group 2, motivating the use of an additional indicator function to differentiate the tails from the main body of their distribution. Fig. 2(a) shows this feature for s_5. Therefore, the regression equation fitted for sales point i is given by

$$\begin{aligned}\mathbb{E}(g_{2i}) = {} & \alpha_0 + \alpha_z I(z_i = 1) + \alpha_1(2) I(x_{1i} = 2) + \ldots + \alpha_5(3) I(x_{5i} = 3) \\ & + \alpha_{1,4}(2,2) I(x_{1i} = 2, x_{4i} = 2) \\ & + \beta_1 I(s_1^{(i)} = 0) + \beta_2 I(s_2^{(i)} = 0) + \beta_5 I(s_5^{(i)} = 0) + \ldots + \beta_9 I(s_9^{(i)} = 0) \\ & + \beta_3 \log(s_3^{(i)}) + \beta_4 \log(s_4^{(i)}) + \gamma_5 I(s_5^{(i)} > 75) + \gamma_7 I(s_7^{(i)} > 75)\end{aligned}$$

where $I(.)$ is the indicator function, $\alpha_k(j)$ is the parameter corresponding to $I(x_{ki} = j)$ and $\alpha_{k,l}(j,h)$ is the parameter corresponding to $I(x_{ki} = j, x_{li} = h)$. See Table 4 for the estimates of the regression parameters, their standard errors and p-values. The resulting model has 23 significant coefficients, with adjusted $R^2 = 0.9929$ and an overall treatment effect of 32.74.

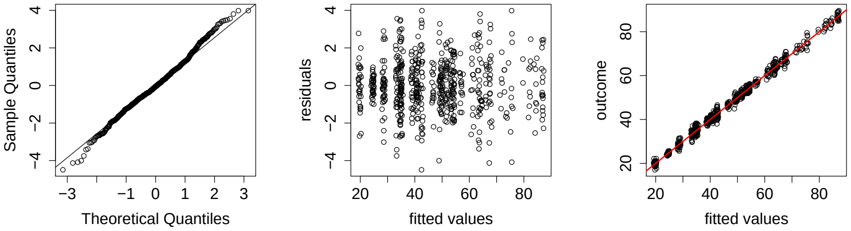


Fig. 4 Residual analysis for group 2: qq-plot of the observed quantiles against the theoretical quantiles, the distribution of the residuals and the observed values against the fitted values of y.

Table 4 Estimates of the regression parameters, their standard errors and associated p-values for group 2.

	Estimate	SE	P-value		Estimate	SE	P-value
α_0	51.79	0.96	< 0.0001	β_1	4.70	1.78	0.008
α_z	32.74	0.17	< 0.0001	β_2	-19.65	3.76	< 0.0001
α_1	-8.46	0.62	< 0.0001	β_3	2.81	0.20	< 0.0001
$\alpha_2(2)$	-1.71	0.58	0.003	β_4	-1.46	0.57	0.01
$\alpha_2(3)$	-5.95	1.07	< 0.0001	β_5	-33.48	2.11	< 0.0001
$\alpha_3(2)$	-19.19	1.17	< 0.0001	β_6	5.85	0.98	< 0.0001
$\alpha_3(3)$	-4.95	1.17	< 0.0001	β_7	25.49	0.58	< 0.0001
$\alpha_3(4)$	-27.05	1.13	< 0.0001	β_8	-10.06	3.73	< 0.0001
α_4	15.86	2.40	< 0.0001	β_9	-37.13	3.63	< 0.0001
$\alpha_5(2)$	5.40	0.86	< 0.0001	γ_5	4.29	0.93	< 0.0001
$\alpha_5(3)$	-6.44	0.39	< 0.0001	γ_7	10.73	0.98	0.005
$\alpha_{1,4}(2,2)$	-11.13	0.84	< 0.0001				

The exploratory analysis performed in Section 2 and the residual analysis shown in Fig. 4 support our proposed model for group 2 that states that, for the sales points that do not sell all the products, the marketing campaign has an average effect of increasing the outcome by €32.74 million.

3.2 Model for Group 1

The covariates s_j for $j = 1, \dots, 9$ are all strictly positive for group 1. However, as discussed in Section 2, the presence of a strong dependence between them encourages the use of PCA to extract from the product sales s_j a subset of orthogonal continuous covariates c_j (see Table 5 for the loadings). Based on the adjusted R^2 increase and standard deviation decrease per number of principal component included in the model, as shown by the plots in Fig. 5, it is apparent that at least the first six principal components should be included in the model, and we find that the first eight principal components give us the best fit.

Table 5 Loadings $l_1, \dots, l_9$ for the PCA by [5].

	l_1	l_2	l_3	l_4	l_5	l_6	l_7	l_8	l_9
$\log(s_1)$	0.371		0.135	0.204	0.239	-0.261	-0.504	-0.609	0.230
$\log(s_2)$	0.310	-0.283	0.265	-0.454	0.269	0.653	-0.200		
$\log(s_3)$	0.165	0.430	-0.125		-0.414	0.107	-0.298	-0.159	-0.684
$\log(s_4)$	0.448	-0.218	-0.796	0.185	0.206		0.155		
$\log(s_5)$	0.397		0.366	0.182	0.350	-0.393		0.502	-0.374
$\log(s_6)$	0.258	0.297	0.292	0.391		0.387	0.602	-0.301	
$\log(s_7)$	0.264	0.340		0.217	-0.286	0.230	-0.350	0.504	0.503
$\log(s_8)$	0.327	0.391		-0.691		-0.300	0.304		0.262
$\log(s_9)$	0.370	-0.572	0.180		-0.671	-0.195	0.114		

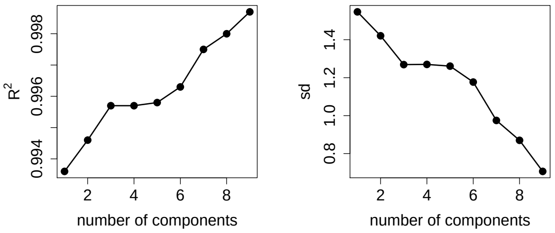


Fig. 5 Adjusted R^2 increase and standard deviation decrease per number of principal component included in the model for group 1.

Following the same arguments regarding multicollinearity as in Section 3.1 and fitting models using a stepwise model selection by AIC to exclude non significant variables and interactions yields the final model which includes: z, x_1, x_2, x_3, x_4, x_5, several interactions terms and the principal components c_k for $k = 1, \ldots, 8$. For sales point i the regression equation is given by

$$
\begin{aligned}
\mathbb{E}(g_{1i}) = {} & \alpha_0 + \alpha_z I(z_i = 1) + \alpha_1 I(x_{1i} = 1) + \ldots + \alpha_5(3) I(x_{5i} = 3) \\
& + \alpha_{1,2}(2,2) I(x_{1i} = 2, x_{2i} = 2) + \ldots + \alpha_{z,5}(1,3) I(z_i = 1, x_{5i} = 3) \\
& + \alpha_{1,2,3}(2,2,2) I(x_{1i} = 2, x_{2i} = 2, x_{3i} = 2) + \ldots \\
& + \alpha_{2,3,4,5}(2,3,1,2) I(x_{2i} = 2, x_{3i} = 3, x_{4i} = 1, x_{5i} = 2) + \gamma_1 c_1^{(i)} + \ldots + \gamma_8 c_8^{(i)}
\end{aligned}
$$

where $I(.)$ is the indicator function, $\alpha_k(j)$ is the parameter corresponding to $I(x_{ki} = j)$ and $\alpha_{k,l}(j,h)$ is the parameter corresponding to $I(x_{ki} = j, x_{li} = h)$ and so on. See Tables 6 and 7 for the estimates of the regression parameters, their standard errors and p-values. The model's fit was checked by the residual analysis shown in Fig. 6.

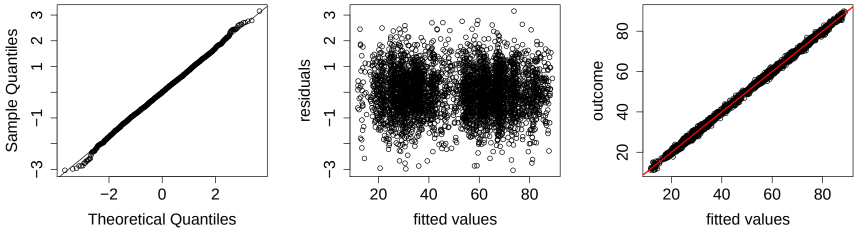


Fig. 6 Residual analysis for group 1: qq-plot of the observed quantiles against the theoretical quantiles, the distribution of the residuals and the observed values against the fitted values of y.

The adjusted R^2 for this model is 0.9980 and the standard error of the outcome for group 1 is 0.8709. The impact of the treatment campaign is significant with an

Table 6 Estimates of the regression parameters, their standard errors and associated p-values for group 1 (part 1, continued in Table 7).

	Estimate	SE	P-value		Estimate	SE	P-value
α_0	32.44	0.09	< 0.0001	$\alpha_{3,4}(4,2)$	0.89	0.59	0.1333
α_z	33.97	0.37	< 0.0001	$\alpha_{3,5}(2,2)$	-8.30	0.54	< 0.0001
α_1	5.23	0.41	< 0.0001	$\alpha_{3,5}(3,2)$	-0.92	0.52	0.0746
$\alpha_2(2)$	7.61	0.26	< 0.0001	$\alpha_{3,5}(4,2)$	0.64	0.58	0.2703
$\alpha_2(3)$	7.83	0.65	< 0.0001	$\alpha_{3,5}(2,3)$	-6.52	0.67	< 0.0001
$\alpha_3(2)$	13.36	0.47	< 0.0001	$\alpha_{3,5}(3,3)$	-0.07	0.56	0.9065
$\alpha_3(3)$	0.76	0.35	0.03	$\alpha_{3,5}(4,3)$	8.60	0.68	< 0.0001
$\alpha_3(4)$	3.11	0.39	< 0.0001	$\alpha_{4,5}(2,2)$	-27.21	0.82	0.9065
α_4	8.90	0.49	< 0.0001	$\alpha_{4,5}(2,3)$	-22.03	0.87	< 0.0001
$\alpha_5(2)$	41.31	0.97	< 0.0001	$\alpha_{z,3}(1,2)$	0.36	0.13	< 0.0001
$\alpha_5(3)$	10.30	0.45	< 0.0001	$\alpha_{z,3}(1,3)$	0.24	0.12	0.05
$\alpha_{1,2}(2,2)$	4.47	0.33	< 0.0001	$\alpha_{z,3}(1,4)$	0.34	0.14	0.01
$\alpha_{1,2}(2,3)$	-2.35	0.66	< 0.0001	$\alpha_{1,2,3}(2,2,2)$	9.89	0.64	< 0.0001
$\alpha_{1,3}(2,2)$	-7.41	0.63	< 0.0001	$\alpha_{1,2,3}(2,3,2)$	16.66	0.72	< 0.0001
$\alpha_{1,3}(2,3)$	-5.79	0.33	< 0.0001	$\alpha_{1,2,3}(2,2,3)$	-12.03	0.76	< 0.0001
$\alpha_{1,3}(2,4)$	-3.97	0.68	< 0.0001	$\alpha_{1,2,3}(2,3,3)$	10.06	0.89	< 0.0001
$\alpha_{1,4}(2,2)$	-6.96	0.38	< 0.0001	$\alpha_{1,2,3}(2,2,4)$	-14.62	0.63	< 0.0001
$\alpha_{1,5}(2,2)$	-20.33	0.81	< 0.0001	$\alpha_{1,2,3}(2,3,4)$	-5.52	0.84	< 0.0001
$\alpha_{1,5}(2,3)$	-1.12	0.65	0.08	$\alpha_{1,2,4}(2,2,2)$	-0.66	0.32	< 0.0001
$\alpha_{2,3}(2,2)$	-22.63	0.62	< 0.0001	$\alpha_{1,2,4}(2,3,2)$	-4.80	0.43	< 0.0001
$\alpha_{2,3}(3,2)$	-16.42	0.62	< 0.0001	$\alpha_{1,2,5}(2,2,2)$	15.68	1.00	< 0.0001
$\alpha_{2,3}(2,3)$	4.95	0.62	< 0.0001	$\alpha_{1,2,5}(2,3,2)$	12.05	0.92	< 0.0001
$\alpha_{2,3}(3,3)$	2.17	0.79	< 0.0001	$\alpha_{1,2,5}(2,2,3)$	-9.14	0.47	< 0.0001
$\alpha_{2,3}(2,4)$	-4.92	0.72	< 0.0001	$\alpha_{1,2,5}(2,3,3)$	-12.35	0.81	< 0.0001
$\alpha_{2,3}(3,4)$	1.08	0.64	0.0926	$\alpha_{2,3,4}(2,2,2)$	-1.01	0.74	0.1709
$\alpha_{2,4}(2,2)$	-2.35	0.51	< 0.0001	$\alpha_{2,3,4}(3,2,2)$	0.83	0.83	0.3190
$\alpha_{2,4}(3,2)$	-1.37	0.55	0.0127	$\alpha_{2,3,4}(2,3,2)$	5.16	0.72	< 0.0001
$\alpha_{2,5}(2,2)$	-49.20	1.17	< 0.0001	$\alpha_{2,3,4}(3,3,2)$	13.00	0.87	< 0.0001
$\alpha_{2,5}(3,2)$	-41.72	1.17	< 0.0001	$\alpha_{2,3,4}(2,4,2)$	9.62	0.79	< 0.0001
$\alpha_{2,5}(2,3)$	-10.99	0.50	0.1333	$\alpha_{2,3,4}(3,4,2)$	6.32	0.73	< 0.0001
$\alpha_{2,5}(3,3)$	-0.30	0.91	< 0.0001	$\alpha_{2,3,5}(2,2,1)$	9.58	0.58	< 0.0001
$\alpha_{3,4}(2,2)$	4.20	0.61	< 0.0001	$\alpha_{2,3,5}(3,2,1)$	8.64	0.64	< 0.0001
$\alpha_{3,4}(3,2)$	-5.64	0.58	< 0.0001	$\alpha_{2,3,5}(2,3,1)$	-6.63	0.59	< 0.0001

average effect of increasing the outcome by around €32.4 million with the presence of the significant interaction term for (z, x_3), causing an increase on the impact of the marketing campaign when this is combined with certain values of the structural variable x_3.

The results presented above, for group 1, are based on the log-transformed product sales. The adjusted R^2 and standard error for the same model with and without the log transformation for the product sales s and outcome y are given in Table 8. This validates our choice, suggested by the plots in Section 2, of taking the logarithm of the product sales while keeping the outcome y on its original scale: this model has the greatest adjusted R^2 and higher precision.

Table 7 Estimates of the regression parameters, their standard errors and associated p-values for group 1 (part 2, continued from Table 6).

	Estimate	SE	P-value		Estimate	SE	P-value
$\alpha_{2,3,5}(3,3,1)$	-6.20	0.65	< 0.0001	$\alpha_{1,2,3,4}(2,2,3,2)$	7.41	1.51	< 0.0001
$\alpha_{2,3,5}(2,4,1)$	0.69	0.67	0.2998	$\alpha_{1,2,3,4}(2,3,3,2)$	6.59	1.59	< 0.0001
$\alpha_{2,3,5}(3,4,1)$	-3.77	0.64	< 0.0001	$\alpha_{1,2,3,4}(2,2,4,2)$	-1.48	1.94	0.44
$\alpha_{2,3,5}(2,2,2)$	7.27	0.68	< 0.0001	$\alpha_{1,2,3,4}(2,3,4,2)$	-17.28	1.47	< 0.0001
$\alpha_{2,3,5}(3,2,2)$	10.28	0.74	< 0.0001	$\alpha_{1,2,3,4}(2,2,2,3)$	16.39	1.66	< 0.0001
$\alpha_{2,3,5}(2,3,2)$	-7.88	0.64	< 0.0001	$\alpha_{1,2,3,4}(2,2,3,3)$	17.36	0.78	< 0.0001
$\alpha_{2,3,5}(3,3,2)$	-2.65	1.07	0.0137	$\alpha_{1,2,3,4}(2,2,4,3)$	2.04	0.82	0.01
$\alpha_{2,3,5}(2,4,2)$	-9.75	0.74	< 0.0001	$\alpha_{1,3,4,5}(2,2,2,2)$	-10.76	1.22	< 0.0001
$\alpha_{2,3,5}(3,4,2)$	-9.68	0.66	< 0.0001	$\alpha_{1,3,4,5}(2,3,2,2)$	-29.91	0.92	< 0.0001
$\alpha_{3,4,5}(2,2,2)$	0.38	0.52	0.4639	$\alpha_{1,3,4,5}(2,4,2,2)$	5.38	0.87	< 0.0001
$\alpha_{3,4,5}(3,2,2)$	6.12	0.50	< 0.0001	$\alpha_{1,3,4,5}(2,2,2,3)$	7.51	1.84	< 0.0001
$\alpha_{3,4,5}(4,2,2)$	0.73	0.50	0.1442	$\alpha_{1,3,4,5}(2,3,2,3)$	-23.15	0.90	< 0.0001
$\alpha_{3,4,5}(2,2,3)$	-2.33	0.60	< 0.0001	$\alpha_{1,3,4,5}(2,4,2,3)$	-7.42	0.98	< 0.0001
$\alpha_{3,4,5}(3,2,3)$	8.88	0.59	< 0.0001	$\alpha_{2,3,4,5}(2,2,2,2)$	-40.27	1.13	< 0.0001
$\alpha_{3,4,5}(4,2,3)$	-2.23	0.54	< 0.0001	$\alpha_{2,3,4,5}(3,2,2,2)$	-4.96	1.08	< 0.0001
$\alpha_{1,3,4}(2,2,2)$	7.64	0.59	< 0.0001	$\alpha_{2,3,4,5}(2,3,2,2)$	-34.48	1.39	< 0.0001
$\alpha_{1,3,4}(2,3,2)$	17.41	0.49	< 0.0001	$\alpha_{2,3,4,5}(3,3,2,2)$	-47.94	1.25	< 0.0001
$\alpha_{1,3,4}(2,4,2)$	8.45	0.54	< 0.0001	$\alpha_{2,3,4,5}(2,4,2,2)$	-40.22	1.48	< 0.0001
$\alpha_{1,3,5}(2,2,2)$	28.75	1.24	< 0.0001	$\alpha_{2,3,4,5}(3,4,2,2)$	-28.73	0.87	< 0.0001
$\alpha_{1,3,5}(2,3,2)$	14.91	0.74	< 0.0001	$\alpha_{2,3,4,5}(2,2,2,3)$	-45.58	0.94	< 0.0001
$\alpha_{1,3,5}(2,4,2)$	20.61	1.30	< 0.0001	$\alpha_{2,3,4,5}(3,3,2,3)$	-15.24	1.02	< 0.0001
$\alpha_{1,3,5}(2,2,3)$	-14.79	1.81	< 0.0001	$\alpha_{2,3,4,5}(2,4,2,3)$	-26.40	0.99	< 0.0001
$\alpha_{1,3,5}(2,3,3)$	5.54	0.62	< 0.0001	$\alpha_{2,3,4,5}(3,4,2,3)$	-12.55	0.90	< 0.0001
$\alpha_{1,3,5}(2,4,3)$	12.50	1.56	< 0.0001	γ_1	0.72	0.03	< 0.0001
$\alpha_{1,4,5}(2,2,2)$	9.80	0.73	< 0.0001	γ_2	-2.28	0.07	< 0.0001
$\alpha_{1,4,5}(2,2,3)$	14.70	0.70	< 0.0001	γ_3	3.69	0.15	< 0.0001
$\alpha_{2,4,5}(2,2,2)$	35.22	1.13	< 0.0001	γ_4	0.64	0.08	< 0.0001
$\alpha_{2,4,5}(3,2,2)$	27.89	0.86	< 0.0001	γ_5	4.60	0.11	< 0.0001
$\alpha_{2,4,5}(2,2,3)$	21.86	0.69	< 0.0001	γ_6	6.35	0.11	< 0.0001
$\alpha_{2,4,5}(3,2,3)$	13.39	0.80	< 0.0001	γ_7	3.50	0.11	< 0.0001
$\alpha_{1,2,3,4}(2,2,2,2)$	-32.64	1.12	< 0.0001	γ_8	-3.93	0.12	< 0.0001
$\alpha_{1,2,3,4}(2,3,2,2)$	-34.55	1.32	< 0.0001				

Table 8 Comparison of several transformations of the outcome y and product sales s for the model for group 1.

	Adjusted R^2	SE
$\log(s)$, $\log(y)$	0.9907	not comparable
$\log(s)$, y	0.9987	0.7062
s, y	0.9941	1.49

4 Forecasting the Outcome for the Seventh Time Period

The second question asked by the organisers of the competition is to forecast the economic return for the seventh time period y_7. In general, a forecast is a statement about the likely course of future events, based on the existing state of knowledge. The most commonly used and natural approach for forecasting of time series is based on the Bayesian paradigm [10]. Statements about the uncertain future are formulated as probabilities conditioned on the available information. However, the first step in a forecasting problem is the construction of a suitable model based on analysis of the known development of the time series. See, for example, [10].

Therefore, we propose here an extension to the regression model we introduced in question one in order to include a time evolution that allows us to forecast.

The organisers of the competition did not specify whether the time periods have constant length, whether they overlap and whether they are strictly sequential, that is, there are no gaps between them. Note that the availability of additional data in this context would have a strong impact on the models we propose below and our predictions. However, in the absence of such information, we assume here that the time periods have constant length, do not overlap and are strictly sequential.

We propose below two models and we use validation tools to choose the most appropriate model for forecasting in our context [7]. We choose to hold out the last available time period, the sixth, for validation. The data which are not held out are used to estimate the parameters of the model, the model is then tested on data in the validation period, and finally forecasts are generated beyond the end of the estimation and validation periods. The outcome y has been observed for the sixth time period, so we can compare the predicted data with the observed data. We will then use the best model selected with the method above to forecast the outcome for the seventh time period. Of course, when we finally forecast the outcome for the seventh time period we use all the available data for estimation, that is, we also use the data available for the sixth time period.

We compare observed data with the predictions for different models by measuring the uncertainty using the Mean Squared Error for the predictions (*MSE*), the Mean Range of the 95% interval for the predictions (*MR*), and the Mean Interval Score (*MIS*). The latter is the mean of the Interval Scores (IS_i). The IS_i for unit i is the score assigned to the prediction interval $(q_{0.025}, q_{0.975})$ when y_{i6} materializes. Reversing the sign of the scoring rule above, we get the negatively oriented interval score,

$$\begin{aligned} IS_i(0.95; y_{i6}) &= (q_{0.975} - q_{0.025}) \\ &\qquad + \frac{2}{\psi}(q_{0.025} - y_{i6}) I(y_{i6} < q_{0.025}) + \frac{2}{\psi}(y_{i6} - q_{0.975}) I(y_{i6} > q_{0.975}), \end{aligned}$$

where $\psi = 0.05$. This score rewards narrow prediction intervals and penalizes (depending on ψ) if the observation misses the interval. The smaller IS_i is, the greater is our forecasting precision.

After describing the validation measures above, in Sections 4.1 and 4.2 below we propose two models and estimate their parameters using the data available for the first five time periods only. The first model that we propose is an autoregressive model AR(1) [4] whereas the second model is a dynamic linear model [10]. A dynamic model allows the inclusion of time-dependent parameters, a realistic assumption in our context. A Bayesian approach was used for both models proposed.

In this approach, in addition to specifying the model $f(y|\theta)$ for the observed data y given a vector of unknown parameters θ, we also assume that θ is a random quantity with prior distribution $p(\theta|\lambda)$, where λ is a vector of hyperparameters to be specified by the modeller. Inference concerning θ is based on its posterior distribution, $p(\theta|y,\lambda)$, that is,

$$p(\theta|y,\lambda) = \frac{f(y|\theta)\pi(\theta|\lambda)}{\int f(y|\theta)\pi(\theta|\lambda)d\theta}. \tag{5}$$

4.1 Autoregressive Model

The first model we consider is a first-order autoregressive model [4], usually referred to as AR(1), where the current outcome y_t depends on the previous outcome y_{t-1} for $t = 2,3,4,5$. This model is motivated by the strong correlation between adjacent outcomes as shown inFig. 2(e)-(f). The AR(1) model, for $t = 2,\ldots,5$ and $i = 1,\ldots,4517$ is given by

$$y_{it} = \phi y_{i,t-1} + v, \quad \text{where} \quad v \sim \mathcal{N}(0,\sigma^2).$$

Moreover, we propose to include in the AR(1) model the marketing campaign z for the first time period, as we have shown in the previous Section its strong impact on outcome for the first time period (see Fig. 2(j)). This gives us

$$y_{i1} = \beta_0 + \beta_1 z_i + v_1, \quad \text{where} \quad v_1 \sim \mathcal{N}(0,\sigma_1^2)$$

with the prior distributions given by

$$\begin{array}{rl} \phi & \sim \text{Gamma}(1,1), \\ \sigma_1^{-2} & \sim \text{Gamma}(0.1,0.01), \\ \sigma^{-2} & \sim \text{Gamma}(0.1,0.01), \end{array} \quad \begin{array}{rl} \beta_0 & \sim \mathcal{N}(30,10), \\ \beta_1 & \sim \mathcal{N}(40,10). \end{array}$$

The integrations required to do inference under (5) for this model are not tractable in closed form, thus we approximated it numerically using a Markov chain Monte Carlo (MCMC) integration algorithm such as Gibbs Sampler [6]. The basic Gibbs Sampler is given by the following. For $\theta = (\theta_1,\ldots,\theta_p)$ and k from 1 to M repeat:

Step 1 Draw $\theta_1^{(k)}$ from $p(\theta_1|\theta_2^{(k-1)},\ldots,\theta_p^{(k-1)},y)$
Step 2 Draw $\theta_2^{(k)}$ from $p(\theta_2|\theta_1^{(k)},\theta_3^{(k-1)},\ldots,\theta_p^{(k-1)},y)$
...
Step p Draw $\theta_p^{(k)}$ from $p(\theta_p|\theta_1^{(k)},\ldots,\theta_{p-1}^{(k)},y)$

Note that for the AR(1) model proposed here $\theta = (\phi,\sigma^2,\sigma_1^2,\beta_0,\beta_1)$. If the conditional distribution is not known in closed form then a Metropolis algorithm is required.

The need for different within-sales point variances for the first time period and the remaining times intervals is confirmed by the results as σ and σ_1 are estimated to be significantly different with $CI(95\%;\sigma) = (12.17, 12.43)$ and $CI(95\%;\sigma_1) = (8.91, 9.31)$. The parameter ϕ that controls the autoregressive evolution over time is estimated by $CI(95\%;\phi) = (0.98, 0.99)$.

4.2 Dynamic Model

The second model we propose is a dynamic linear mixed model which combines sales point-specific random effects with explanatory variables whose coefficients may vary over time.

In the context of the first question we replaced the product sales by their log-transformations. We refrain from performing a transformation in this case as it would confront us with a mixture model and a complex fitting procedure for dynamic models. The natural choice would be to include the product sales or their corresponding principal components. We investigated these options but such models gave poor fitting. However, modeling of group 2 in question 1 revealed that the use of indicator functions such as $I(s_{jt} > 75)$ is relevant for the description of the outcome of interest.

Fig. 2(k)-(l) show the association between outcome and the structural variables over time. It appears that only the effect of x_1 on y_t varies as time passes, therefore we choose the corresponding parameter of x_1 to vary over time while the parameters of the remaining structural variables remain constant. We also decided to include autoregressive sales point-specific random effects α_{it} in order to quantify the variation between different sales points. This is an important component of the model as it aims to capture the effect of unit specific variables that we did not include in the model.

In summary, we propose the following model:

$$\begin{aligned} y_{it} &= \beta_{0t} + \beta_{1t} z_i + \beta_{2t} I(s^{(i)}_{2t} > 75) + \beta_{3t} I(s^{(i)}_{3t} > 75) + \beta_{4t} I(s^{(i)}_{5t} > 75) \\ &\quad + \beta_{5t} I(s^{(i)}_{6t} > 75) + \delta_{1t} I(x_{1i} = 2) + \delta X_i + \alpha_{it} + v_t, \end{aligned} \tag{6}$$

where

$$\begin{aligned} v_t | \sigma_t^2 &\sim \mathcal{N}(0, \sigma_t^2), & \delta_{1t} | \delta_{1,t-1} &\sim \mathcal{N}(\delta_{1,t-1}, w^2_{\delta_1}) \\ \alpha_{it} | \alpha_{i,t-1}, w^2 &\sim \mathcal{N}(\alpha_{i,t-1}, w^2), & \beta_{kt} | \beta_{k,t-1} &\sim \mathcal{N}(\beta_{k,t-1}, w_k^2) \text{ for } k = \{0, 1, \ldots, 5\}, \end{aligned}$$

with $\sigma_1^2 \neq \sigma_2^2 = \ldots = \sigma_6^2 = \sigma^2$, $\delta = (\delta_2, \ldots, \delta_9)'$ and $X_i = I(x_{2i} = 2), I(x_{2i} = 3), I(x_{3i} = 2), I(x_{3i} = 3), I(x_{3i} = 4), I(x_{4i} = 2), I(x_{5i} = 2), I(x_{5i} = 3)'$ for $t = 2, \ldots, 5$ and $i = 1, \ldots, 4517$. Furthermore, $w_{\delta_1} = w_k = 1$ for $k = \{0, 1, \ldots, 5\}$. The prior distributions used are given by

model	MSE	MIS	MR
1	145.50	57.00	48.14
2	119.88	51.85	43.10

Table 9 MSE, MIS and MR for the validation dataset and the prediction of y_6.

$$\begin{aligned} w^{-2} &\sim \text{Gamma}(1,1), \\ \sigma_1^{-2} &\sim \text{Gamma}(0.1,0.01), \\ \sigma^{-2} &\sim \text{Gamma}(0.1,0.01), \\ \beta_{01} &\sim \mathcal{N}(40,1), \\ \beta_{11} &\sim \mathcal{N}(30,1), \end{aligned} \qquad \begin{aligned} \beta_{k1} &\sim \mathcal{N}(0,1), \text{ for } k=2,3,4,5, \\ \alpha_{i1} &\sim \mathcal{N}(0,w^2), \\ \delta_k &\sim \mathcal{N}(0,10), \text{ for } k=2,\ldots,9, \\ \delta_{11} &\sim \mathcal{N}(0,1). \end{aligned}$$

Samples may be generated from the model described using a Markov chain Monte Carlo (MCMC) algorithm [6] as described in Section 4.1.

The assumption of different within-sales point variances in model 2 is confirmed by the model fitting as the standard deviations of the response, for the model obtained by integrating out the sales point specific random term, are

$$CI(95\%; \sqrt{\sigma_1^2 + w^2}) = (7.35, 7.57)$$

and

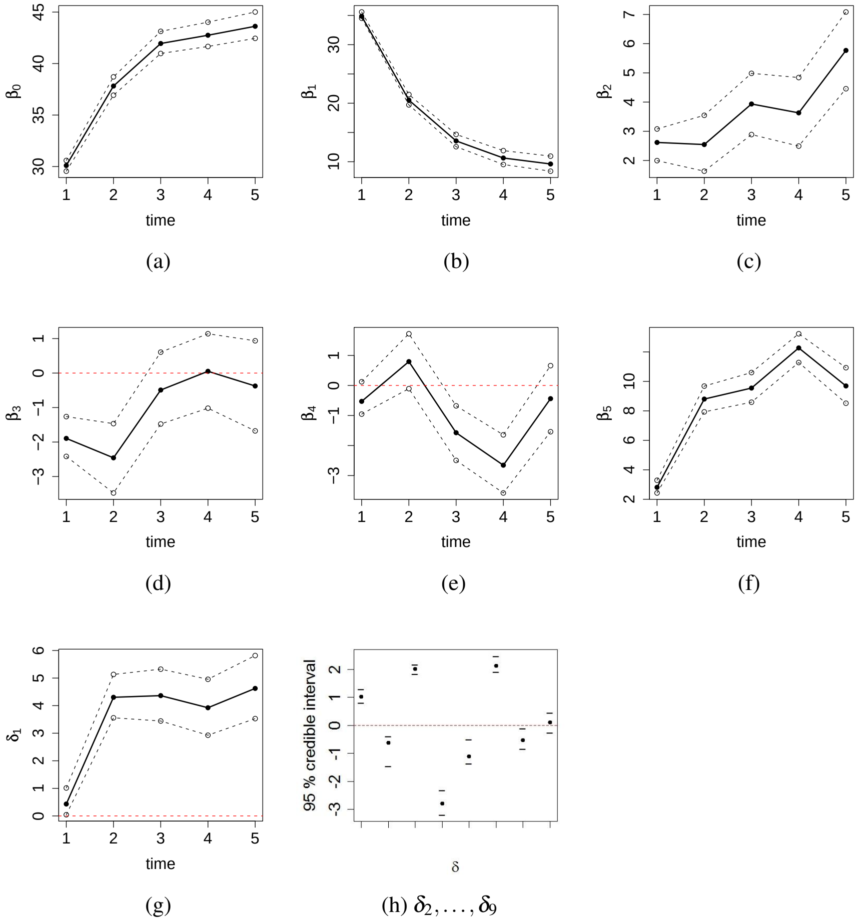
$$CI(95\%; \sqrt{\sigma^2 + w^2}) = (9.47, 9.80).$$

Moreover, the results show that the dynamics in the coefficients are very important for some of the parameters, as shown in Fig. 7. We observe that the overall mean is monotonically increasing over time (Fig. 7(a)), whereas the treatment effect decreases as time passes (Fig. 7(b)). This is in line with the time evolution shown in Fig. 2(j).

Moreover, it appears that the parameters associated with the indicators of the tails for the product sales s_{2t}, s_{3t}, s_{5t} and s_{6t}, shown in Fig. 7(c)-(f), vary substantially over time. Note the larger credible intervals due to the small number of sales point with extreme values. The time varying effect of x_i on the outcome, observed in Fig. 2(k), is also confirmed by Fig. 7(g). Furthermore, the coefficients that do not vary over time (δ_k; $k = 2, \ldots, 9$) are significant except for δ_9, corresponding to the effect of $I(x_{5i} = 3)$.

4.3 Validation Results

We now compare observed data with the predictions obtained for the sixth time period from the two models proposed by measuring their uncertainty and precision. The validation measures for both models are summarized in Table 9 and we observe that model 2 is superior to model 1 using any of the three criteria.



(a) (b) (c)

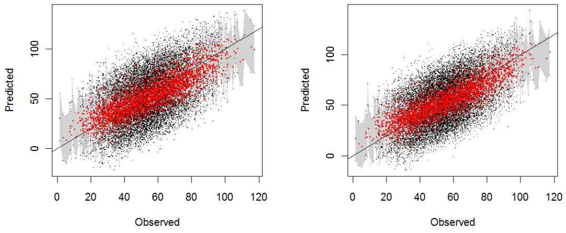
(d) (e) (f)

(g) (h) $\delta_2, \ldots, \delta_9$

Fig. 7 Median (solid line) and 95% credible intervals (dashed line)for the mean parameters involved in model (6) using the validation dataset, that is, the data for the first five time periods.

Furthermore, Fig. 8(a) shows that the predicted values are reasonable, and that the range of the 95% credible intervals for the predictions obtained for the autoregressive model is larger than the range obtained for the dynamic model.

From this validation analysis we conclude that the dynamic model gives better predictions and represents well the variability of the data. Thus, we proceed to predict data for the seventh time period using all the available data provided by the organisers.



(a) AR(1) model (b) Dynamic model

Fig. 8 Plots of the predicted values against the corresponding observed outcomes y_{i6} for the first model, AR(1), and the dynamic model fitted, respectively. The red dots represent the mean of the predicted values and the smaller black dots represent the limits of the credible interval. The interval is highlighted with grey.

4.4 Predictions

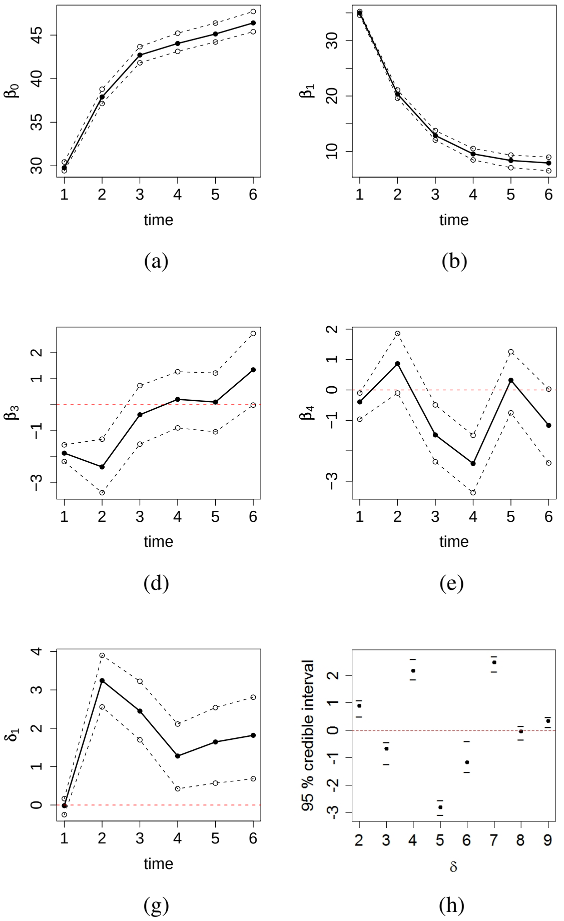
Following our results above, we fit the observation equation (6) to predict the economic return during the seventh time period using all the data provided by the organisers. Thus, taking into account the uncertainty in the estimation of all time-dependent parameters, i.e. w_k^2 $(k = 0,...,5)$ and $w_{\delta_1}^2$, but also the within-sales point variance σ^2 and the between-sales point variance w^2 the distribution for the predictions is given by:

$$y_{i7}|\beta_{06},\ldots,\beta_{56},\delta_{16},\delta,\alpha_{i6},\sigma^2,w^2 \sim \mathscr{N}(\mu_{i7},\sigma_{i7}^2)$$

with

$$\begin{aligned}\mu_{i7} &= \beta_{06}+\beta_{16}z_i+\beta_{26}I(s_{27}^{(i)}>75)+\beta_{36}I(s_{37}^{(i)}>75)\\&\quad+\beta_{46}I(s_{57}^{(i)}>75)+\beta_{56}I(s_{67}^{(i)}>75)+\delta_{16}I(x_{1i}=2)+\delta X_i+\alpha_{i6},\\ \sigma_{i7}^2 &= w_0^2+w_1^2 z_i+w_2^2 I(s_{27}^{(i)}>75)+w_3^2 I(s_{37}^{(i)}>75)+w_4^2 I(s_{57}^{(i)}>75)\\&\quad+w_5^2 I(s_{67}^{(i)}>75)+w_{\delta_1}^2 I(x_i=2)+\sigma^2+w^2.\end{aligned}$$

Fitting the model in equation (6) for the first six time periods and drawing from the above distributions yields the predictions summarized in Fig. 9. Note that the model captures very different behaviors in the evolution of the parameters over time. The overall mean (β_0), shown in Fig. 9(a), increases over time while the effect of treatment (β_1) decreases (Fig. 9(b)). In addition, the intervals for β_1 are very narrow suggesting that the effect of treatment on outcome is precisely estimated. The variable $I(s_{2t}>75)$ seems to be very influential on the outcome with an effect that increases over time as we see in Fig. 9(c). The effects of $I(s_{3t}>75)$ is also increasing with time but it is non-significantly different from zero for $t = 3,4,5$ (Fig. 9(d)).



(a) (b) (c)

(d) (e) (f)

(g) (h)

Fig. 9 95% credible intervals (dashed line) and median (solid line) for the mean parameters of model (6) using all the data available $y_1, \ldots, y_6$.

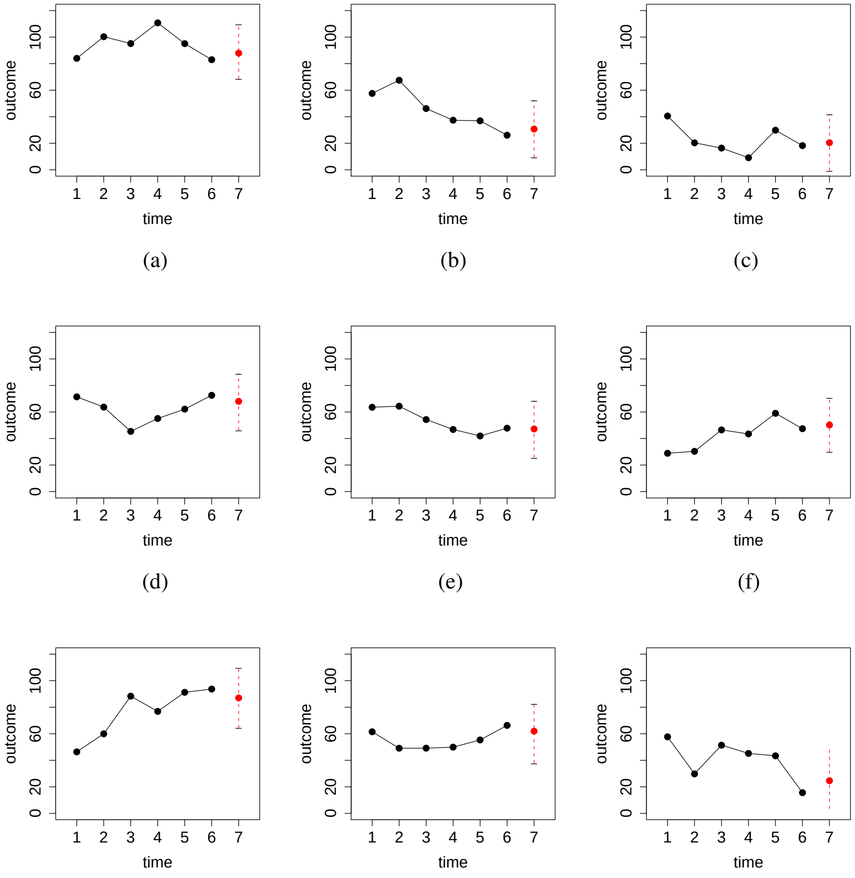
The effect of $I(s_{5t} > 75)$ and $I(s_{6t} > 75)$ on outcome seem non-linear (Fig. 9(e)-(f). The uncertainty in the estimation of the effect of x_1 over time is rather large, but the effect seems to be positive for $t > 1$ as we can see in Fig. 9(g). Fig. 9(h) shows the effect of the other structural variables that are all significant except for one level of x_5 ($x_{5i} = 2$). The standard deviations for the model obtained by integrating out the sales point specific random term are

$$CI(95\%; \sqrt{\sigma_1^2 + w^2}) = (7.15, 7.37)$$

and

$$CI(95\%;\sqrt{\sigma^2+w^2}) = (9.47, 9.80).$$

Finally, we predict the outcome for the seventh time period and we include plots for the time series of the outcome for a selection of the sales points in Fig. 10. Note the different shapes of such time series, and how the dynamic model that we propose captures the variability. Also, see in Fig. 11 the prediction of the seventh time period for some values of the structural variables.

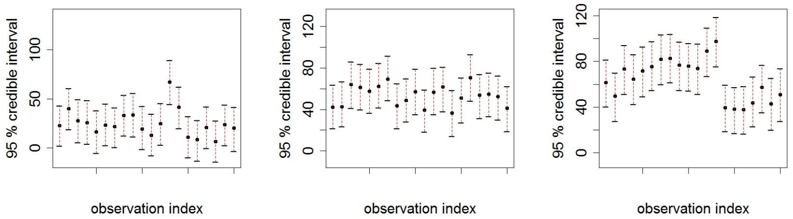


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 10 Time series for y with the prediction for the seventh time period with its credible interval for sales points 1 (a), 1058 (b), 1102 (c), 1204 (d), 1851 (e), 2168 (f), 3112 (g), 848 (h), 862 (i).



(a) $x_1 = 1$ (b) $x_1 = 2$ (c) $x_1 = 2, z = 1$

Fig. 11 Prediction of y_7 for some of the sales points with $x_1 = 1$ (a), $x_1 = 2$ (b) and $x_1 = 2$, $z = 1$ (c).

5 Discussion

We have analysed the data provided by the organisers of the competition and proposed an approach to answer the two questions.

For the first question, that required the evaluation of the impact of the treatment campaign on the outcome, we proposed a mixture of regression models. Our results show a good fit and the residual analysis confirmed the plausibility of our assumptions. Moreover, our model was able to extrapolate the additional effect of the marketing campaign when it is associated with certain values of the other available covariates, that is, many interaction terms were highly significant. We have presented here an extensive analysis of the relationship between the outcome variable and the other covariates for the first time period.

The second question asked by the organisers of the competition is to forecast the economic return for the seventh time period. We performed this task within the Bayesian paradigm and proposed autoregressive and dynamic models. We tested our model on its prediction of the sixth time period using several validation measures. Our tests showed that the Bayesian dynamic model was the most accurate while still capturing the variability of the data and avoiding overfitting. Hence, we selected this model to predict the economic return for the seventh time period, the task required by the competition. We believe that our Bayesian dynamic model incorporates the main features of the data provided but it is also easy to adapt to the arrival of new information in real time, by updating the prior distributions or by including new covariates.

Acknowledgments

We thank the organisers of the First Cladag Data Mining Prize for their constructive comments that helped improve this paper and the SAS Institute for sponsoring the prize. Moreover, we thank CRiSM (Centre for Research in Statistical Methodology, Department of Statistics, University of Warwick) for financial support.

References

1. Belsley, D.A., E. Kuh, and R.E. Welsch (2004). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity, 1st ed., John Wiley & Sons, Inc. New York.
2. Benzecri, J. (1992). Correspondence analysis handbook, vol. 125 of Statistics: Textbooks and Monographs.
3. Chambers, J. M. (1992). Linear models. Chapter 4 of Statistical Models in S, eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
4. Chatfield, C. (2003). The Analysis of Time Series: an Introduction. CRC Pr I Llc.
5. Croux, C., Filzmoser, P. and Oliveira, M. (2007). Algorithms for Projection-Pursuit Robust Principal Component Analysis. Chemometr. Intell. Lab. , Vol. 87, pp. 218-225.
6. Gamerman, D. and Lopes, H.F. (2006). Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Chapman & Hall/CRC.
7. Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. Journal of the American Statistical Association, 102(477): pp.360–378.
8. Hastie, T. J. and Pregibon, D. (1992). Generalized linear models. Chapter 6 of Statistical Models in S eds J. M. Chambers and T. J. Hastie, Wadsworth & Brooks/Cole.
9. O'Brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality and Quantity 41(5):673–690.
10. Pole, A. and West, M. and Harrison, J. (1994). Applied Bayesian forecasting and times series analysis. Chapman & Hall/CRC.