THE 2020 ELECTION IN THE UNITED STATES: BETA REGRESSION VERSUS REGRESSION QUANTILES

Jan Kalina

Abstract

The results of the presidential election in the United States in 2020 desire a detailed statistical analysis by advanced statistical tools, as they were much different from the majority of available prognoses as well as from the presented opinion polls. We perform regression modeling for explaining the election results by means of three demographic predictors for individual 50 states: weekly attendance at religious services, percentage of Afroamerican population, and population density. We compare the performance of beta regression with linear regression, while beta regression performs only slightly better in terms of predicting the response. Because the United States population is very heterogeneous and the regression models are heteroscedastic, we focus on regression quantiles in the linear regression model. Particularly, we develop an original quintile regression map; such graphical visualization allows to perform an interesting interpretation of the effect of the demographic predictors on the election outcome on the level of individual states.

Key words: elections results, electoral demography, quantile regression, heteroscedasticity, outliers

JEL Code: D72, C21, Y91

Introduction

Understanding the demographic characteristics of voters should be an important step in a detailed analysis of any presidential, parliamentary, gubernatorial, or municipal elections. This paper is interested in a unique analysis of the results of presidential election in the United States of America in the year 2020 by means of advanced statistical methods. The presidential election took place in November 2020 with Donald J. Trump as the incumbent president desiring to be re-elected, and Joseph (Joe) R. Biden as his main challenger, who subsequently became the 46th president of the United States. While opinion polls presented by a variety of different providers predicted almost unanimously Joe Biden to win with a strong lead before Donald Trump, the final results are well known to have been surprisingly tight.

Political scientists around the world are naturally interested in a detailed analysis of results of elections based on demographic data (Gerring et al., 2015). The results of the previous U.S. presidential election of 2016 were intensively discussed from the point of view of political science, demography, economics, psychology, psychohistory, or other disciplines (Einhorn, 2018). Jones et al. (2017) claimed that voters express a certain direct message to the politicians through the elections; in the USA, discovering this message may be even more important than finding the actual winner of the presidential election. Analyzing results of elections (not only in the context of the United States) represents a useful source of demographic knowledge about the population in a given country, as discussed in Tavares et al. (2020). All these references confirm that a statistical analysis of the results of elections in general may be useful for several reasons, which also include the possibility to obtain a broader perspective of the future political development.

While simple analyses by exploratory (descriptive) tools of the 2020 U.S. presidential elections have been presented in abundant forms in media, this paper desires to analyze the results by more advanced tools. Regression quantiles, currently very popular in econometrics (Koenker, 2017), appear to find applications also in political science. For example, Okada (2018) used regression quantiles to find infant mortality rate and life expectancy to be positively impacted in democratic countries compared to dictatorships; regression quantiles were selected as a suitable tool due to the non-normal distribution of the demographic characteristics under consideration. Regression quantiles still represent an object of theoretical research. Regression quantiles within statistical functionals were studied by Jurečková et al. (2020), who focused on properties of averaged regression quantiles with the ability to mask the influence of regressors (predictors, independent variables). Also regularized versions of regression quantiles are gaining on importance; these are suitable for correlated regressors, removing the necessity to perform a dimensionality reduction. Still, we can say that standard as well as regularized regression quantiles are not robust with respect to outliers, as recalled in the overview of Kalina (2013) of robust methods suitable for data mining.

Regression quantiles or beta regression, where the latter represents a generalized linear model suitable for data with a response in the form of percentages, seem to have been only rarely used for the analysis of results of elections. Section 1 of the current paper describes the data used in our computations. Beta regression is used in Section 2, linear regression in Section 3, and regression quantiles for the linear regression model in Section 4.

1 Data description

In this paper, we consider the results of the presidential election as well as demographic characteristics as values over the individual $n = 50$ states of the USA. District of Columbia is not considered in our analysis, as it is extremely specific in all there predictors as well as in the response. Let us describe the response as well as three predictors corresponding to selected demographic characteristics about individual U.S. states:

- $Y = (Y_1, \dots, Y_n)^T$ as the response variable corresponds to the percentage of votes for Donald J. Trump[1], the incumbent president.
- $X_1 = (X_{11}, \dots, X_{n1})^T$ represents the weekly church attendance[2], defined as the percentage in the state population of those who attend a church, synagogue or mosque once a week or almost every week, as estimated in 2015;
- $X_2 = (X_{12}, \dots, X_{n2})^T$ represents the percentage of Afroamerican population[3] in the state population in 2015;
- $X_3 = (X_{13}, \dots, X_{n3})^T$ represents the population density[4] (as the number of inhabitants per square kilometer) in 2015.

The response as well as all predictors are continuous variables. These three predictors seem to play a crucial role in predicting the election results in a given individual state, i.e. in modeling the response conditioned on fixed values of the predictors, where the latter are able to distinguish between a rural conservative state on one hand and a urban liberal (or socialist) state with a strong support for the Black Lives Matter movement on the other hand.

All computations in this paper were performed in R software (R Core Team, 2017). The map of Figure 1 shows Y across the 50 states, where darker shades corresponds to a higher percentage of electoral vote for Donald Trump. Maps of the three predictors are shown as left images of Figures 2 to 4, where darker shades correspond to a higher weekly church attendance, higher percentage of Afroamerican population, and larger population density in a given state, respectively. Such maps were created using the library usmap of R software.

Interpreting the given data should naturally start with an exploratory data analysis; numerous elementary results of the analysis were however presented in the media. Before proceding to more advanced statistical methods, let us only very briefly mention that the

[1] https://en.wikipedia.org/wiki/2020_United_States_presidential_election

[2] https://en.wikipedia.org/wiki/Church_attendance

[3] https://en.wikipedia.org/wiki/Demographics_of_the_United_States

[4] https://en.wikipedia.org/wiki/List_of_states_and_territories_of_the_United_States_by_population_density

predictors are mutually correlated. Particularly, Pearson correlation coefficient r evaluated for pairs of predictors are $r(X_1, X_2) = 0.57$, $r(X_1, X_3) = -0.21$, and $r(X_2, X_3) = 0.22$.

Tab. 1: Mean square error evaluated for different regression models $\boldsymbol{Y \sim X_1 + X_2 + X_3}$.

Regression method	MSE
Beta regression	46.09
Least squares in (1)	46.98
L1-estimator in (1)	49.66

Source: own computations

2 Beta regression

Beta regression represents a generalized linear model especially suitable for a response variable in the form of percentages. Usually, it is claimed that linear regression is not be very suitable for such proportional data, especially under heteroscedasticity. Nevertheless, it is not clear in a particular data analysis task, whether a linear model would be sufficient or not; in fact, the literature presenting results of beta regression only rarely performs comparisons with results of linear regression. Cribari-Neto & Zeileis (2010) described the computation of beta regression in R software using library betareg. Pereira (2019) investigated a suitable transformation of residuals of beta regression ensuring their approximate standard normal distribution. In addition, regression quantiles have been available for beta regression (Lu & Fan, 2020), but we are not aware of their publicly available implementation.

Let us overview results of the beta regression model, where Y is used as the response of three predictors X_1, X_2, and X_3. All three predictors are statistically significant using the Wald test on the level $\alpha = 0.05$; X_2 has the largest p-value (only 0.013), while the p-value of X_1 is about 10^{-7} and that of X_3 is 0.004. The pseudo-R^2 in the model is 0.55. The mean square error, i.e. the common measure of prediction ability of (not only) the beta regression. is presented in Table 1. There, values of the response were taken as percentages so that e.g. 50 corresponds to 0.50.

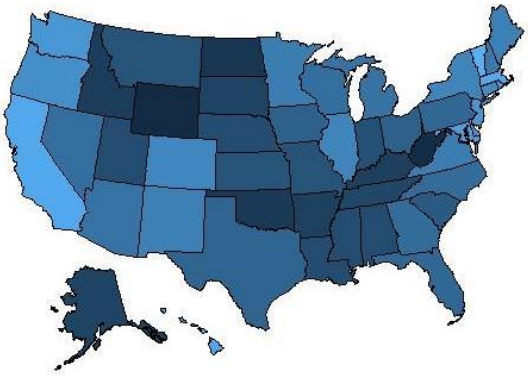
3 Linear regression

We also consider the linear regression model in the form

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + e_i, \qquad i = 1, \dots, n. \tag{1}$$

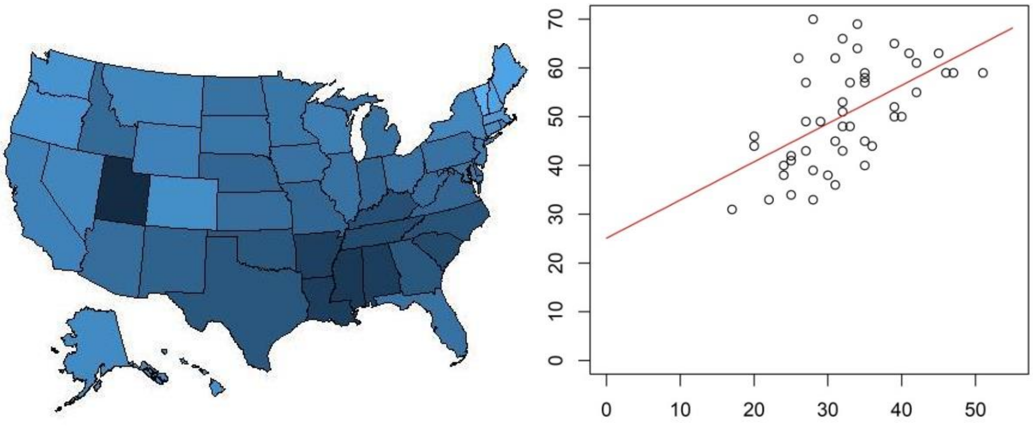
An inspection of scatter plots of the response against individual predictors (as shown in the right images of Figures 2 to 4) reveals that Y depends on X_1 and X_3 in models, which can be more or less approximated by normal models with strong heteroscedasticity.

Fig. 1: Map of the United States shaded according to Y, i.e. percentage of popular vote in each state for Donald J. Trump.



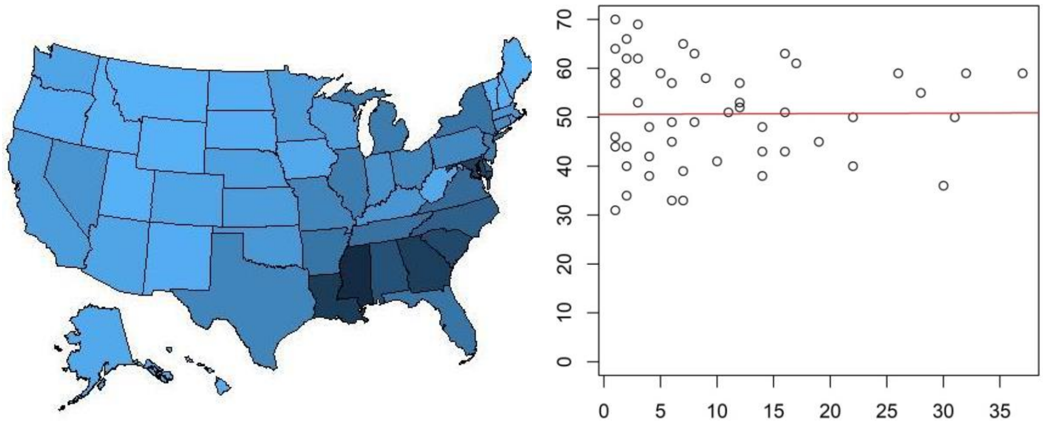
Source: own graph using the data from wikipedia (cited in Section 1)

Fig. 2. Left: Map shaded according to X_1. Right: Y against X_1 (with a linear trend estimated by least squares).



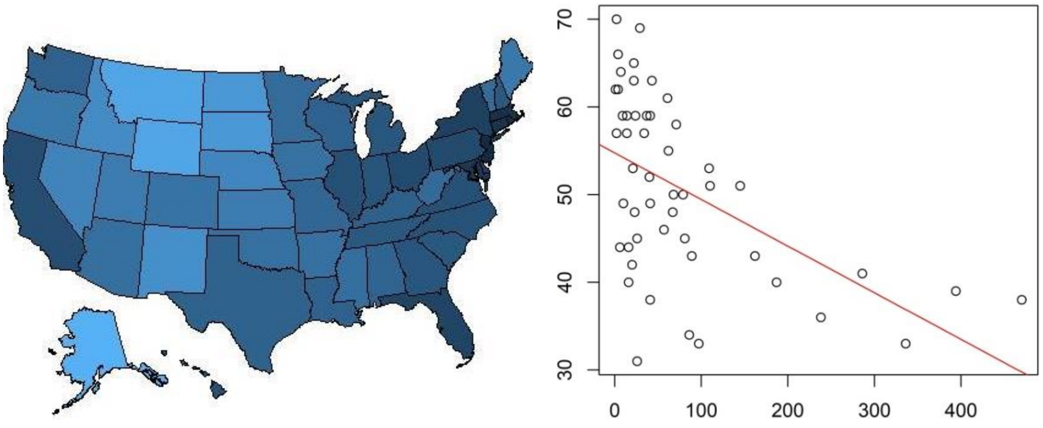
Source: own graph using the data from wikipedia (cited in Section 1)

Fig. 3. Left: Map shaded according to X_2. Right: Y against X_2 (with a linear trend estimated by least squares).



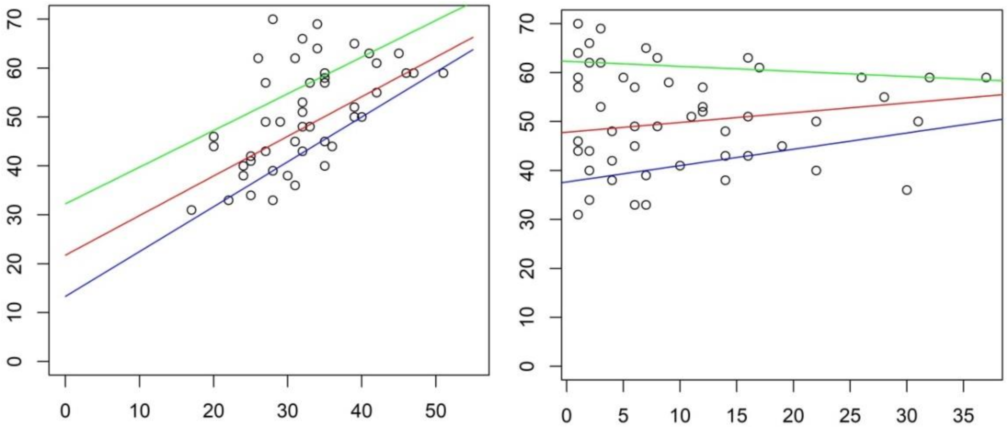
Source: own graph using the data from wikipedia (cited in Section 1)

Fig. 4. Left: Map shaded according to X_3. Right: Y against X_3 (with a linear trend estimated by least squares).



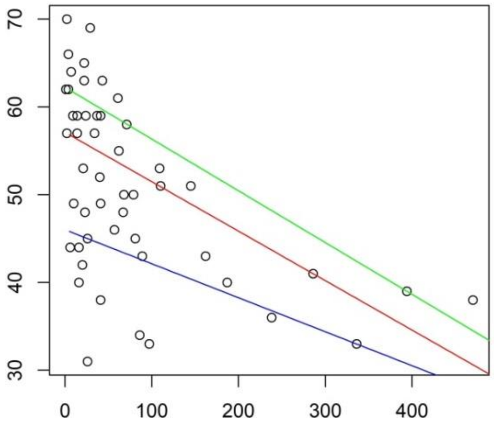
Source: own graph using the data from wikipedia (cited in Section 1)

Fig. 5. Left: Regression quantiles in the model $Y{\sim}X_1$. Right: Regression quantiles in the model $Y{\sim}X_2$.



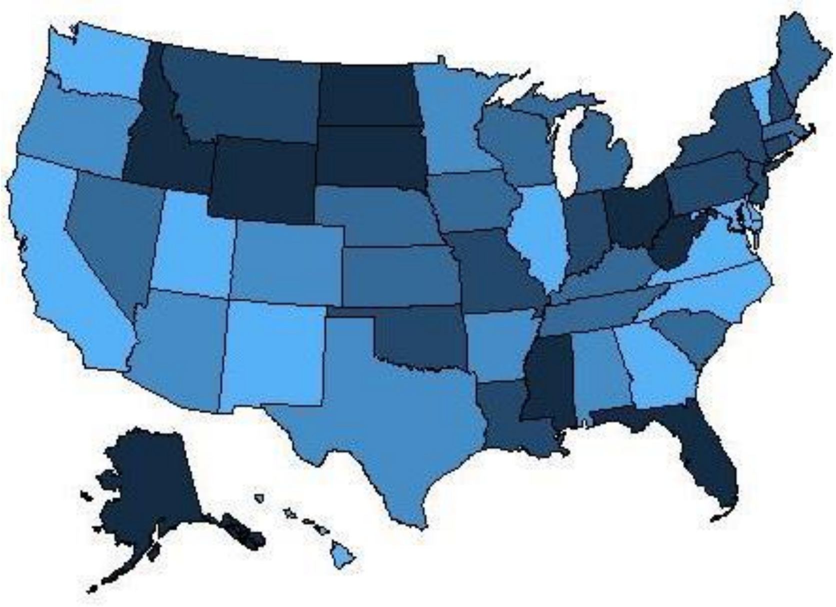
Source: own computations

Fig. 6: Regression quantiles in the model $Y \sim X_3$.



Source: own computations

Fig. 7: Regression quintile map of the United States of Section 4.1, shaded according to the quintiles of residuals in (1) for each individual state.



Source: own computations

The response Y does not seem to depend on X_2. However, when X_2 is omitted from the model, R^2 drops to 0.50; therefore, we decided to keep X_2 in the model, as it contributes to the variability of the response especially through the strong correlation with X_1. The coefficient of determination in the linear regression model (1) is equal to $R^2 = 0.55,$ i.e. the same as the pseudo-R^2 in the beta regression. As Table 1 reveals, beta regression is slightly better than the least squares estimator in (1) in terms of the mean square error.

4 Regression quantiles

We still assume the linear regression model (1) and use a set of 4 regression quantiles (i.e. quintiles) to get a richer information about the relationship of Y on the three predictors. Figures 5 and 6 reveal regression quantiles in models, where Y is always explained by one of the three predictors. The images show the 0.25 quantile (blue), 0.50 quantile (red), and the 0.75 quantile (green). The regression quantiles were computed using the library quantreg of R software. We continue with the analysis, although the 0.50 quantile (i.e. the regression median) is not particularly good in explaining the response, as revealed in Table 1; it remains less suitable compared to the beta regression or to the least squares in (1).

The images in Figures 5 and 6 considers only one of the three available regressors. Still, it is interesting to see that the quantiles in the model $Y{\sim}X_1$ are almost parallel. The model $Y{\sim}X_2$ however without any apparent trend of the response and especially $Y{\sim}X_3$ shows a clear heteroscedasticity. Let us pay a closer attention to regression quantiles in the particular fit of Y depending only on X_3. There are several (statistically) influential states with a leverage effect in the regression model. The response Y in these states (New Jersey, Rhode Island, Massachusetts, Connecticut, and Maryland) is much below the nation-wide mean.

4.1 Regression quintile map

Finally, we consider an original type of map exploiting the information contained in the regression quantiles. For the computation of this map denoted as a "regression quintile map", the regression quantiles with $\tau = 0.2, \tau = 0.4, \tau = 0.6,$ and $\tau = 0.8$ were computed. The corresponding 4 regression lines divide all the 50 states to 5 categories (i.e. below the first quintile; between the first and the second; etc.), which are depicted in Figure 7. There, the lightest shade of the blue color corresponds to the lowest quintile (with e.g. California or Illinois) and the darkest to the highest quintile (with e.g. Idaho or Mississippi).

The regression quintile map allows to conclude interesting results for individual states. We can say the very intensive campaign of Donald Trump in Florida and Ohio was evidently unnecessary, as his outcomes in both states are in the highest quintile. On the other hand, his feeble campaign in Arizona was underrated, as he was defeated there by Joe Biden; Trumps's outcome there is in the second quintile. Still, the most important battleground states where Donald Trump lost by a narrow margin were Pennsylvania and Georgia. Our model reveals Donald Trump to score relatively well (i.e. with regard to the three predictors) in Pennsylvania, as his outcome is in the fourth quintile there. On the other hand, the result in

Georgia is only in the lowest quintile, although Georgia is a conservative rural state of the Bible Belt traditionally (but not this time) representing a Republican bastion.

Conclusions

Especially the idea of the regression quintile map of Section 4.1 seems appealing to bring a new knowledge about the results of the presidential election in the USA in 2020. The map (see Figure 7) is a simple tool comprehensible for demographers or political scientists, is able to capture local effects on the level of individual states and present them comprehensibly. These local effects may be considered in predictions of the outcome before future elections. Still, the result of the 2020 election was not much different from the results of 2016; the electoral stability as a measure of aggregate stability of a democratic system was discussed in Mainwaring et al. (2017). Our analysis with only three predictors is of course simplistic. Political science journals present in fact much more detailed results from past elections around the world (i.e. on the level of counties), as presented e.g. by Gerring et al. (2015). We do not model turnout here and neither do we compare the results with results of Republican primaries in 2016, which again reflected local effects on the level of individual states.

From the statistical point of view, this work mainly reveals the potential of regression quantiles. While beta regression is usually recommended if the response contains percentages, it offers slightly better predictions in terms of MSE compared to linear estimators here. However, there seem no regression quantiles for beta regression to be available in R software. The proposed regression quintile map is meaningful if applied to continuous predictors. We actually use regression quantiles (with a suitable plot) for regression diagnostics and for interpretations on the level of individual observations, which seems as a unique approach as well. Observations (i.e. states) in an extreme quantile can be interpreted as outliers, although regression quantiles have actually been only rarely exploited for outlier detection.

As future research, we plan to consider also other regression estimators, including robust tools of Saleh et al. (2012) suitable for models with measurement errors, or more complex types of regression quantiles. The latter include nonlinear (for a specified nonlinear model) or nonparametric regression quantiles, or quantile regression neural networks. We verified these nonlinear quantiles to be suitable only for larger datasets, while the dataset with the 50 states does not seem sufficiently large for this purpose.

Acknowledgment

The work was supported by the project GA21-19311S of the Czech Science Foundation.

References

Cribari-Neto F., Zeileis A. (2010). Beta regression in R. Journal of Statistical Software, 34(2), 1-24.

Einhorn J. (2018). Election 2016: A psychohistorical, psychoeconomic analysis of the 2016 United States Presidential Election. Journal of Psychohistory, 45, 192-211.

Gerring J., Palmer M., Teorell J., Zarecki D. (2015). Demography and democracy: A global, district-level analysis of electoral contestation. American Political Science Review, 109, 574-591.

Jones J.J., Bond R.M., Bakshy E., Eckles D., Fowler J.H. (2017). Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 U.S. presidential election. PLOS One, 12(4), e0173851.

Jurečková J., Picek J., Schindler M. (2020). Empirical regression quantile processes. Applications of Mathematics, 65, 257-269.

Kalina, J. (2013): Highly robust methods in data mining. Serbian Journal of Management, 8, 9-24.

Kalina J., Vašaničová P., Litavcová E. (2019). Regression quantiles under heteroscedasticity and multicollinearity: Analysis of travel and tourism competitiveness. Ekonomický časopis, 67, 69-85.

Koenker R. (2017). Quantile regression: 40 years on. Annual Review of Economics, 9, 155-176.

Lu X. & Fan Z. (2020). Generalized linear mixed quantile regression with panel data. PLoS ONE 15(8), e0237326.

Mainwaring S., Gervasoni C., España-Najera A. (2017). Extra- and within-system electoral volatility. Party politics 23, 623-635.

Okada K. (2018). Health and political regimes: Evidence from quantile regression. Economic Systems, 42, 307-319.

Pereira G.H.A. (2019). On quantile residuals in beta regression. Communications in Statistics—Simulation and Computation, 48, 302-316.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Saleh A., Picek J. & Kalina J. (2012). R-estimation of the parameters of a multiple regression model with measurement errors. Metrika, 75, 311-328.

Tavares A.F., Raudla R. & Silva T. (2020). Best of both worlds? Independent lists and voter turnout in local elections. Journal of Urban Affairs, 42, 955-974.

Contact

Jan Kalina
The Czech Academy of Sciences, Institute of Computer Science
Pod Vodárenskou věží 2, 182 07, Prague 8, Czech Republic
& Charles University, Faculty of Mathematics and Physics
Sokolovská 83, 186 75 Prague 8, Czech Republic
kalina@cs.cas.cz