Spatial Statistics Project

Topic: Agricultural yield estimation of German administrative districts for which official yield data is not provided. For this study, I mainly benefited from Fahrmeir et al.'s book; and I used Fleet and Jepson's handout for summary of Markov random fields. These sources are given below:

Fahrmeir, L., Kneib, T., Lang, S., & Marx, B. (2013). *Regression: Models, Methods and Applications.* Berlin Heidelberg: Springer-Verlag.

Fleet, D., & Jepson, A. *Markov Random Fields.* Retrieved July 08, 2017, from http://www.cs.toronto.edu/~kyros/courses/2503/Handouts/mrf.pdf

1. Introduction

In this study, I am trying to estimate agricultural yield of German administrative districts for which official yield data is not provided in the Regional Database Germany (www.regionalstatistik.de). This database provides detailed official statistics for various subjects however there are plenty of missing data in the provided statistics. To estimate the missing data I will use a spatial smoothing method which involves Markov random field and penalized least square criterion.

2. Background

2.1. Nonparametric regression

In some real life applications, as in this study, independent variables have nonlinear effects on dependent variables. It is often difficult to model these nonlinear effects with usual parametric approaches. Nonparametric regression models allow us to estimate nonlinear effects in a flexible way. More explicitly, nonparametric methods do not require any restrictive assumption such as finding a certain parametric functional form fitting the data. For example, in the case of one covariate x, the standard model for nonparametric regression is defined as $y_i = f(x_i) + \varepsilon_i$. For the error variable ε_i, the same assumptions as in the simple linear regression model apply. The function f is estimated in a data-driven way through nonparametric approaches. Despite their flexibility in modelling the data, nonparametric approaches can be considered as a way of exploratory data analysis which helps researchers to find simpler parametric functional forms.

2.2. Spatial smoothing

Spatial smoothing is a method used for nonparametric regression. In general, spatial smoothing can be exercised for two different problems: (1) spatial effects of location variables measured on a continuous scale, (2) spatial effects of discrete spatial units. And discrete spatial units can be represented in form of regions (i.e. districts of a country) or sections on a discrete grid. However, these two different problems can overlap in some cases. For example, if we have a large number of discrete spatial units, we can use a continuous spatial model to analyze such data.

In the case of continuous spatial information (1), the Euclidean distance can be used to find the distance between two locations. Whereas in the discrete spatial information (2), we need a different concept to describe the spatial arrangement of the data, and this concept relies on proper mapping of neighborhoods. Neighborhoods can be defined in several ways; two of them are given below:

- In case of regions, the spatial covariate s is a member of a particular region s. This membership is usually defined by common boundaries. Modification of this definition may be needed if some regions are islands and/or observation area is divided into separate locations.

- In case of grids, usually the nearest four or eight neighbors are used.

2.3. Markov random field

I will use spatial smoothing to estimate spatial effects of discrete spatial units represented by regions. To do this, I need to define the relationship between regions, and Markov random field (MRF) will help me for this definition. A MRF is a graph (Fleet & Jepson) $G = (V, E)$ such that,

- $V = \{1,2, \dots, d\}$ is the set of nodes, each of which is associated with a random variable, γ_s, for $s = 1,2, \dots, d$
- The neighborhood of node s, denoted by N_s, is the set of nodes to which s is adjacent; i.e. $r \in N_s$ if and only if $(s, r) \in E$
- and MRF satisfies $p(\gamma_s \mid \{\gamma_r\}_{r \in V \setminus s}) = p(\gamma_s \mid \{\gamma_r\}_{r \in N_s})$.

The key property of MRFs is the transmission of information to a long distance on the graph through local connections. This communication between nodes of the graph constructs the basis for spatial smoothing.

2.4. Penalized least square

Penalized least square criterion (PLS) will be used to estimate the spatial effects of regions.

- The notation $s \sim r$ denotes that region s and r are neighbors.
- Every region has its own regression coefficient $f_{geo}(s) = \gamma_s$, $s = 1,2, \dots, d$.
- To have smooth spatial effects, regression coefficients of nearby regions should not divert too much from each other. A penalty criterion based on squared differences of regression coefficients can be used for nearby regions.
- PLS criterion: $PLS(\lambda) = \sum_{i=1}^{n} (y_i - f_{geo}(s_i))^2 + \lambda \sum_{s=2}^{d} \sum_{r \in N(s), r<s} (\gamma_r - \gamma_s)^2$
- In the PLS criterion, $N(s)$ defines the neighbors of region s.

The penalty term contains all possible combinations of neighboring regions and each combination is considered only once. This penalty discourages large deviations in regression coefficients of nearby regions.

2.5. The model and solution

- The model: $y_i = f_{geo}(s_i) + \varepsilon_i$ such that $\varepsilon \sim N(0,\ \sigma^2)$
- The model in matrix notation: $y = Z\gamma + \varepsilon$
- The design matrix Z: $Z[i, s] = \begin{cases} 1 & if\ y_i\ was\ observed\ in\ region\ s\ and \\ 0 & otherwise. \end{cases}$
- The vector of function evaluations: $f_{geo} = Z\gamma$
- The penalty term: $\lambda \gamma' K \gamma$
- The penalty matrix: $K[s, r] = \begin{cases} -1 & s \neq r,\ s \sim r, \\ 0 & s \neq r,\ s \nsim r, \\ |N(s)| & s = r. \end{cases}$
- By minimizing the PLS criterion, the PLS estimate is obtained: $\hat{\gamma} = (Z'Z + \lambda K)^{-1} Z' y$

The PLS estimate $\hat{\gamma}$ is the vector of spatial effects for each region, in other words, it is the final product that we would like to have after spatial smoothing. We should keep in mind that this estimate can change for different values of smoothing parameter λ.

2.6. Concerns about this spatial smoothing method

The main concern about the method explained above is its simplicity. The PLS estimate $\hat{\gamma}$ covers all possible covariate effects in itself, that is, effects of other important non-spatial covariates would be merged into spatial effects. Therefore, simultaneous consideration of non-spatial and spatial effects is necessary to get improved estimates.

In chapter 9 of the Regression book (Fahrmeir, Kneib, Lang, & Marx, 2013), more advanced techniques considering all kinds of effects together are explained. These techniques are based on mixed models and Bayesian approaches. Chapter 9.8 of Fahrmeir et al.'s book (Case Study: Malnutrition in Zambia) is quite useful to understand these subjects; and it employs "BayesX" software for estimating models.

3. Agricultural yield estimation

I used R software for this study. As a first step to spatial data analysis, state based potato yield (100kg/ha) of Germany is plotted (see Figure 1) on a map from Global Administrative Areas database (www.gadm.org). In the figure, the map on the left is colored by user codes; and from black to white potato yield increases. As seen in this map, there are only three states in red color for which official data is not provided. However, this number will increase when we look at the district based potato yield of Germany. (In Figure 1, the map on the right is colored by ready R functions; but they did not work well in showing states without yield observation)

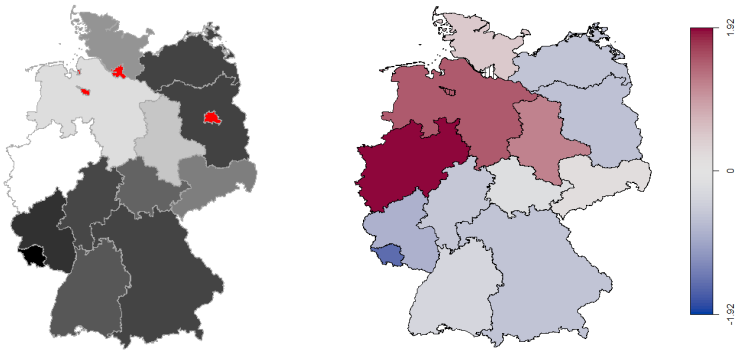

Figure 1 - State based potato yield of Germany

District based potato yield of Germany and the estimated spatial effects for different values of smoothing parameter λ are plotted in Figure 2. As seen in the first map, we have yield observation for a limited number of districts (230 out of 439). To estimate the spatial effects of German districts on potato yield, I followed the steps given below. After the estimation, we see that smoothness of the estimate increases as λ increases, however residual variance also increases with λ.

- Penalty matrix K is calculated by observing the neighborhood structure of German districts.
- Ruegen is the only district which consists solely of islands. So, penalty matrix K is updated by defining a neighbor to Ruegen. Otherwise, the matrix in the solution becomes singular.
- Design matrix Z is computed according to districts of each observation.
- Model is estimated with penalized least square criterion. Several values are used for smoothing parameter λ.

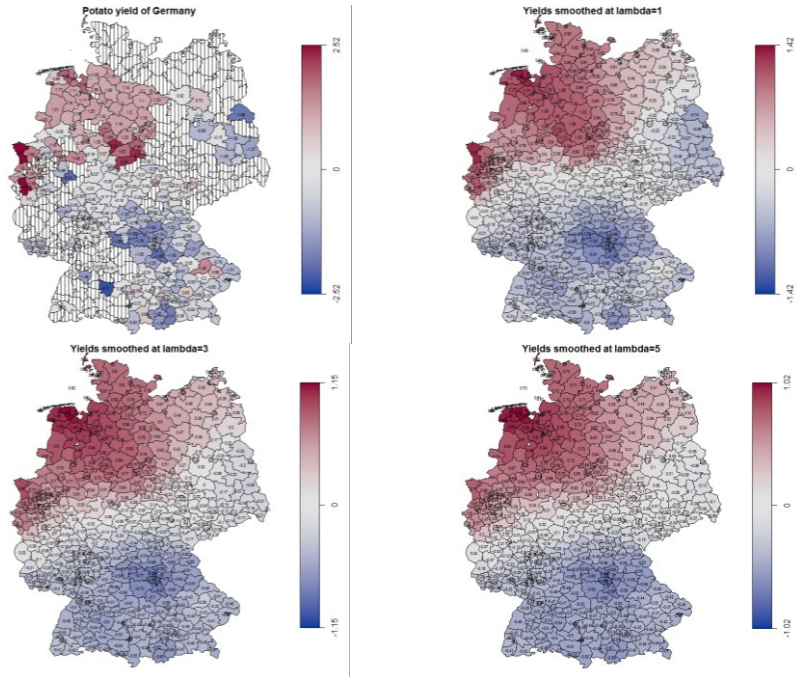

Figure 2 - District based potato yield of Germany and estimated spatial effects

As the last step, assumptions are tested. Residuals reveal homoscedastic and linear behavior (see Figure 3). And residuals are normally distributed (see Figure 4).

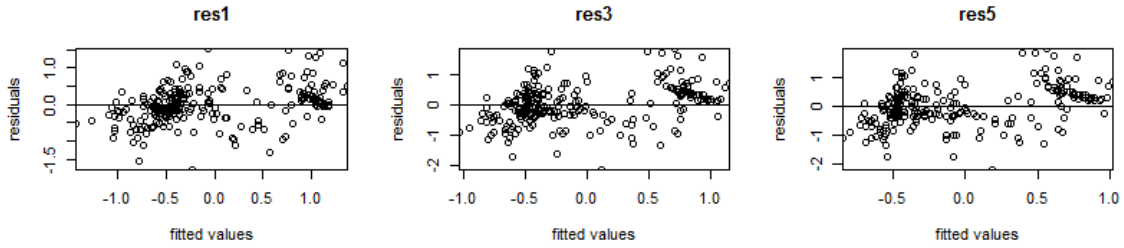

Figure 3 - Homoscedastic and linear residuals (especially when lambda=1)

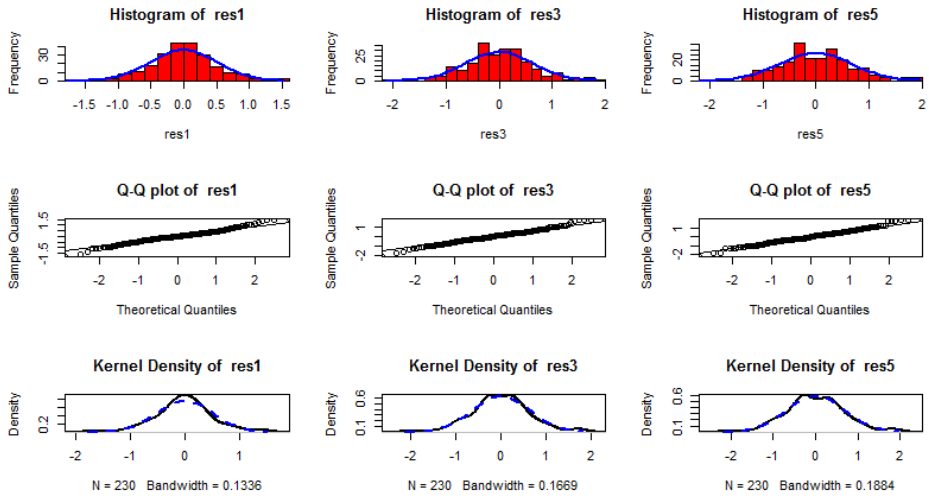

Figure 4 - Normally distributed residuals