Stat 155

Rima Izem

Areal Data Analysis

Review...

- Examples in Areal Data: German UR, SID in NC, SAT scores in US, census data in Boston-Brookline.
- Spatial information for Areal data: Adjacency matrix.
- Adjacency matrix, two ingredients

Review...

- Examples in Areal Data: German UR, SID in NC, SAT scores in US, census data in Boston-Brookline.
- Spatial information for Areal data: Adjacency matrix.
- Adjacency matrix, two ingredients proximity and strength

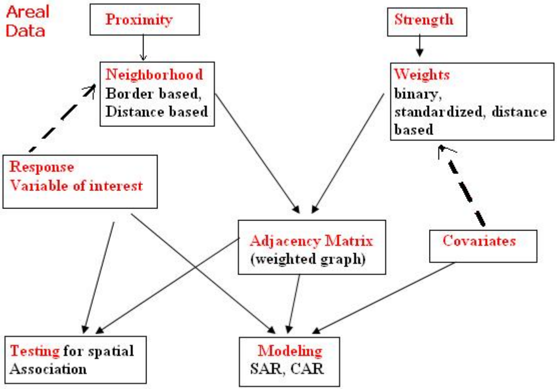
Areal data, what and why

- Areal data, other name: Lattice data.
- Response of interest Y_i measured in block or **areal unit** B_i.
- Areal models of spatial variation (CAR and SAR), goal not so much interpolation as accounting for spatial pattern in linear model and/or spatial smoothing.

Areal data, general outline

- Representation of spatial proximity in areal data using *weighted* graphs
- Testing for spatial pattern: Global testing using Moran's I or Geary's C statistic
- Modelling spatial pattern using SAR or CAR.
- special topics, responses are counts, space-time...etc

Areal data, general outline



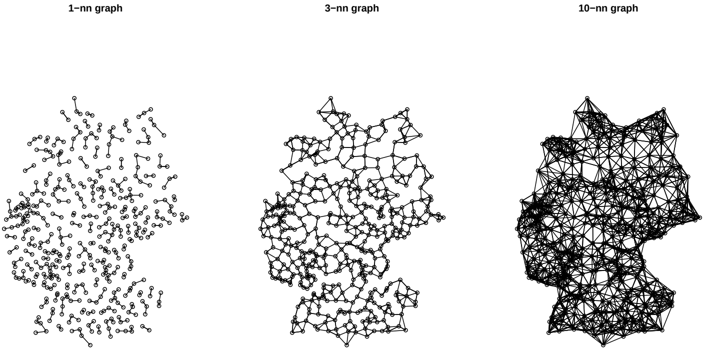
Proximity

- Border based. Two areal units are neighbors if they share a border.
- Distance based.
  - k-Nearest neighborhood, where the neighborhood of an areal unit is its k *nearest* areal units.
  - ϵ neighborhood, two areal units are neighbors if their centroids are within a *distance* ϵ of each other.

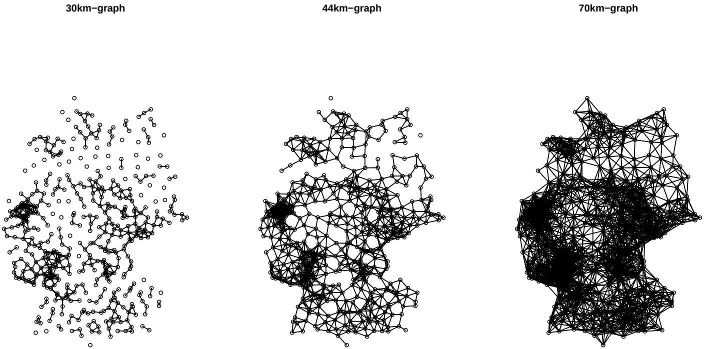
Distance based proximity

- Centroids of areal data: geographic point of mass, or largest city, or political center
- Distance: Euclidean distance (or driving distance or driving time..etc) between centroids; mean driving distance, mean driving time, walking distance...etc
- Choice of k or ϵ,
  - Model constraint: connected graph
  - Sanity check: *reasonable* and problem specific

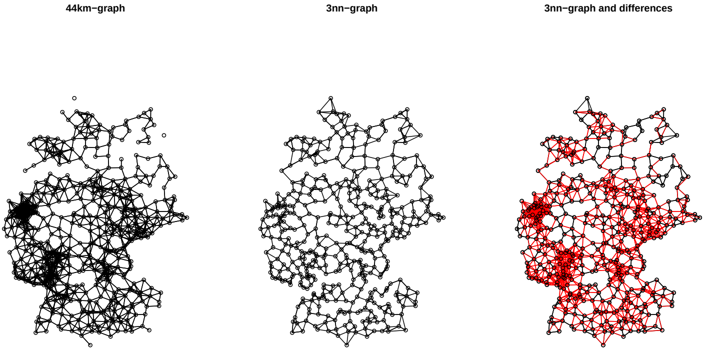
Germany data, choice of k?



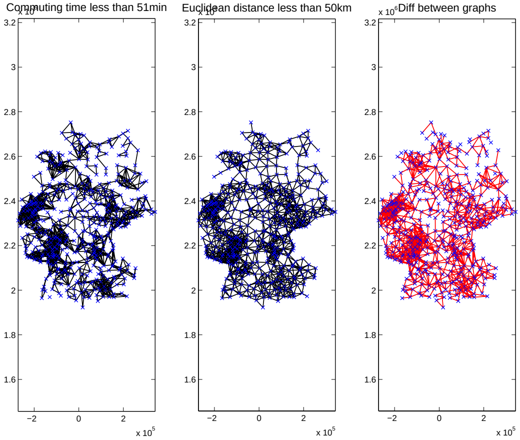
Germany data, choice of ϵ?



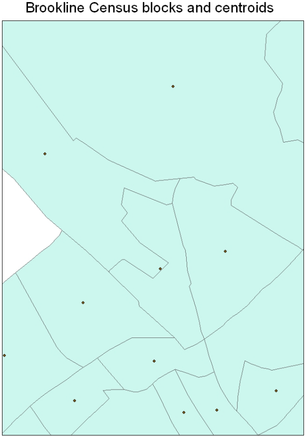
Differences between k and ϵ



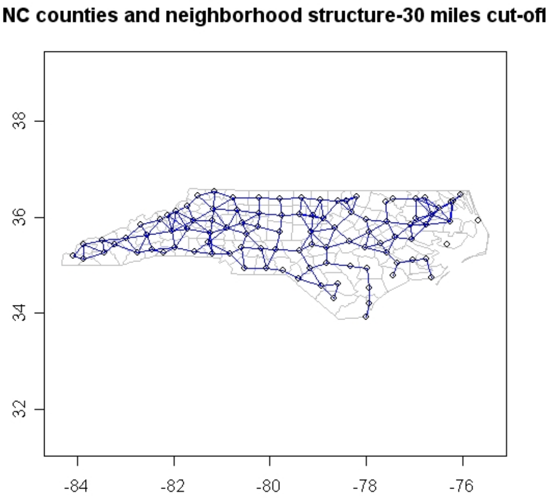
Germany data, choice of distance?



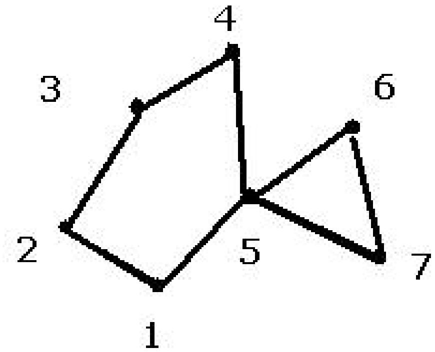
Beware of using centroids with irregular shape blocks



SID data, proximity



Adjacency Matrix, Toy Example



Toy Example, Neighborhood information

1	2	5		
2	1	3		
3	2	4		
4	3	5		
5	1	4	6	7
6	5	7		
7	6	5		

Toy Example, Adjacency Matrix

Binary matrix

$$\begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}$$

Toy Example, Adjacency Matrix

row standardized

$$
\begin{pmatrix}
0 & 0.5 & 0 & 0 & 0.5 & 0 & 0 \\
0.5 & 0 & 0.5 & 0 & 0 & 0 & 0 \\
0 & 0.5 & 0 & 0.5 & 0 & 0 & 0 \\
0 & 0 & 0.5 & 0 & 0.5 & 0 & 0 \\
0.25 & 0 & 0 & 0.25 & 0 & 0.25 & 0.25 \\
0 & 0 & 0 & 0 & 0.5 & 0 & 0.5 \\
0 & 0 & 0 & 0 & 0.5 & 0.5 & 0
\end{pmatrix}
$$

Areal, adjacency Matrix

Adjacency Matrix is a matrix of neighbor's weights.

- Binary weights (denoted by style B in spdep in R), $b_{ij} = 1$ if unit j is neighbors of unit i, $b_{ij} = 0$ otherwise.
- Standardized weights
  - Standardized by number of neighbors, or row standardized $w_{ij} = \frac{b_{ij}}{b_{i+}}$ if unit j is neighbor of unit i, $w_{ij} = 0$ otherwise, where b_{i+} is the row sum ($=$ the number of neighbors)
  - Standardized by number of links, $c_{ij} = \frac{b_{ij}}{b_{i+}+b_{+j}}$ if unit j is neighbor of unit i, $c_{ij} = 0$ otherwise

Areal, adjacency Matrix (contd)

we can also use distance between units and some covariates (ex: population size) to determine strength of relationship

- Ex: $w_{ij} = \frac{1}{d_{i,j}}$ if unit j is neighbor of unit i
- Ex2: In North Carolina SID data (Cressie, 1992) $w_{ij} = \frac{min\{d_{ij}: i=1,\ldots,n\}}{d_{ij}} (\frac{n_j}{n_i})^{0.5}$ where n_i is population size of unit i.
- Note that adjacency matrix need not be symmetric. (ex: sids weights, or nearest neighbor weights).

Testing spatial association

Measuring strength of association and testing for spatial association,

- Moran's I
- Geary's C

Strength of association, Moran's I

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{(\sum_{j \neq i} w_{ij}) \sum_i (Y_i - \bar{Y})^2}$$

- Null hypothesis: no-spatial association, i.e Y_i's are i.i.d
- Under null hypothesis, $\frac{I + 1/(n-1)}{\sqrt{Var(I)}} \approx N(0, 1)$. (asymptotic result)
- Hypothesis testing: Use asymptotic normality or permutation test to find p-value

Example, Moran's I output for German's data

3-nn graph, binary weights

```
        Moran's I test under randomisation

data:  URdata[, 4]
weights: GermWeight3nneigh

Moran I statistic standard deviate = 23.4337, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.852422419      -0.002283105       0.001330309
```

Example, Moran's I output for German's data

70km graph, binary weights

```
        Moran's I test under randomisation

data:  URdata[, 4]
weights: GermWeight70km

Moran I statistic standard deviate = 49.0027, p-value < 2.2e-16
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     0.8246345675     -0.0022831050      0.0002847641
```

Strength of association, Geary's C

$$\frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{\sum_{i \neq j} w_{ij}(Y_i - \bar{Y})^2}$$

- Null hypothesis: no-spatial association, i.e Y_i's are i.i.d
- Under the null hypothesis, $\frac{C-1}{\sqrt{Var(C)}} \approx N(0,1)$.
- Hypothesis testing: Use asymptotic normality or permutation test to find p-value

Example, Geary's C output for German's data

3-nn graph, binary weights

```
        Geary's C test under randomisation

data:  URdata[, 4]
weights: GermWeight3nneigh

Geary C statistic standard deviate = -22.4626, p-value < 2.2e-16
alternative hypothesis: less
sample estimates:
Geary C statistic       Expectation          Variance
      0.137356052       1.000000000       0.001474827
```

Example, Geary's C output for German's data

70km graph, binary weights

```
        Geary's C test under randomisation

data:  URdata[, 4]
weights: GermWeight

Geary C statistic standard deviate = -46.6458, p-value < 2.2e-16
alternative hypothesis: less
sample estimates:
Geary C statistic       Expectation          Variance
     0.1772549729      1.0000000000      0.0003111033
```

Local Moran's I

Recall, global Moran's :

$$I = \frac{n \sum_i \sum_j w_{ij}(Y_i - \bar{Y})(Y_j - \bar{Y})}{\left(\sum_{j \neq i} w_{ij}\right) \sum_i (Y_i - \bar{Y})^2}$$

Local Moran's I, at point i, let

$$I_i = \frac{n(Y_i - \bar{Y}) \sum_{j \text{ neighbor of } i} w_{i,j}(Y_j - \bar{Y})}{\sum_k (Y_k - \bar{Y})^2}$$

Use normality assumption on each I_i and find z scores (high $|z|$ score is evidence of clustering)

Local G and G^*

Recall, global Geary's C

$$\frac{(n-1)\sum_i \sum_j w_{ij}(Y_i - Y_j)^2}{\sum_{i \neq j} w_{ij}(Y_i - \bar{Y})^2}$$

Local G score, at point i

$$G_i = \frac{\sum_{j \text{ neighbor of } i} Y_j}{\text{number of neighbors of } i}$$

Local G^* score at point i, counts the point itself as a neighbor

$$G_i^* = \frac{Y_i + \sum_{j \text{ neighbor of } i} Y_j}{1 + \text{ number of neighbors of } i}$$

Summary

- Global test, see if there is any evidence for spatial association
- LISA: localized indicators of Spatial Autocorrelation: (local Moran's I or local G) are exploratory test for clustering in the data
- When using localized test, control for multiple testing (using Bonferroni, or FDR)

Local Moran's

Example: Local Moran's at one location

```
> myLocalINoCorr[436:439,]
            Ii         E.Ii    Var.Ii      Z.Ii   Pr(z > 0)
[1,] 1.0360022 -0.002283105 0.06295603  4.138072 1.751181e-05
[2,] 0.8050364 -0.002283105 0.13766464  2.175877 1.478224e-02
[3,] 1.2580829 -0.002283105 0.11362455  3.739044 9.236065e-05
[4,] 2.7577576 -0.002283105 0.06755979 10.618700 1.219671e-26
> myLocalI[436:439,]
            Ii         E.Ii    Var.Ii      Z.Ii   Pr(z > 0)
[1,] 1.0360022 -0.002283105 0.06295603  4.138072 3.327245e-04
[2,] 0.8050364 -0.002283105 0.13766464  2.175877 1.478224e-01
[3,] 1.2580829 -0.002283105 0.11362455  3.739044 1.293049e-03
[4,] 2.7577576 -0.002283105 0.06755979 10.618700 1.951474e-25
```

multiple testing and correction

Recall, under the null hypothesis (no spatial association),

- $I_i \sim N(E(I_i), Var(I_i))$, equivalent to z_i score follows a standard normal where $z_i = \frac{I_i - E(I_i)}{\sqrt{Var(I_i)}}$
- p-value (two sided) is $P(N(0,1) > |z_i| or N(0,1) < -|z_i|)$.

multiple testing and correction

- Hypothesis test procedure: decide on a significance level (say 5%), reject the null hypothesis if p-value $< 5\%$
- Meaning of p-value?

multiple testing and correction

- Hypothesis test procedure: decide on a significance level (say 5%), reject the null hypothesis if p-value $< 5\%$
- Meaning of p-value?

$$pvalue = P(\text{rejecting } H_0 | H_0 \text{ is true })$$

- So, even if there is no clustering in the data, if you perform 1000 tests, you might have 5% of the test be significant, just by chance.

From one test to multiple tests

If you are making 2 simultaneous tests, then you have a composite null hypothesis. So, H_0 is: there is no spatial association at location 1 and no spatial association at location 2, i.e.

$$H_0 = H_{0,1} \text{ and } H_{0,2}$$

where $H_{0,i}$ is that there is no spatial association at location i.

- Bonferroni correction is based on the following inequality

$$\text{pvalue}_{\text{simultaneous}} < \text{pvalue}_1 + \text{pvalue}_2$$

- **So, using Bonferroni's correction, to have global significance level of 5%, use a 2.5% (=global significance/(number of tests)) significance level for each individual test**

multiple testing and correction

Bonferroni's correction in spdep package for multiple tests considers that the number of multiple tests for each region is only taken as the number of neighbours + 1 for each region, rather than the total number of regions.

SAR model

SAR: simultaneous autoregressive model, model for exponential family distribution

- Gaussian with mean zero
- Autoregressive regression (lag model and SAR model)
- General model (poisson example)

SAR model, contd

$$
\begin{aligned}
Y_i &= \sum_j c_{i,j} Y_j + \epsilon_i \\
\text{Observed Values} &= \text{Spatial Signal} \; + \; \text{independent residuals}
\end{aligned}
$$

Observed value is an average of *neighboring* observations (hence the *auto* in autoregression).

SAR model, contd

$$
\begin{aligned}
Y_i &= \sum_j c_{i,j} Y_j + \epsilon_i \\
\text{Observed Values} &= \text{Spatial Signal} + \text{independent residuals}
\end{aligned}
$$

Observed value is an average of *neighboring* observations (hence the *auto* in autoregression). Two questions: IS this model well defined?

SAR model, contd

$$
\begin{aligned}
Y_i &= \sum_j c_{i,j} Y_j + \epsilon_i \\
\text{Observed Values} &= \text{Spatial Signal} + \text{independent residuals}
\end{aligned}
$$

Observed value is an average of *neighboring* observations (hence the *auto* in autoregression). Two questions: IS this model well defined? How is matrix C related to Adjacency matrix W?

SAR model, gaussian case

Individual specification (local):

$$\begin{aligned} Y_i &= \sum_j c_{i,j} Y_j + \epsilon_i \\ \text{Observed Values} &= \text{Spatial Signal} + \text{independent residuals} \end{aligned}$$

or equivalently, in Matrix form (global)

$$(I - C)Y = \epsilon$$

where $\epsilon \sim MN(0, D)$, and $D = diag(\sigma_1^2, \ldots, \sigma_n^2)$. Is this model well defined? (i.e. does this model define a valid multivariate distribution)?

SAR model, gaussian case (contd)

In model

$$(I - C)Y = \epsilon, \text{ where } \epsilon \sim N(0, D)$$

ϵ induces the following distribution for Y

$$Y \sim N(0, (I - C)^{-1} D((I - C)^{-1})')$$

if and only if $(I - C)$ is **full rank.**

SAR model, gaussian case (contd)

Common choice for C: $C = \lambda W$

- λ is called the spatial autoregression parameter.
- Model becomes

$$Y_i = \lambda \sum_{j \text{ neighbor } i} w_{ij} Y_i + \epsilon_i$$

SAR model, gaussian case (contd)

Equivalent choice for C: $C = \alpha \tilde{W}$, where $\tilde{W}$ is the weighted adjacency matrix.

- α is called the spatial autocorrelation parameter.
- Model becomes

$$Y_i = \alpha \sum_{j \text{ neighbor } i} \frac{w_{ij}}{\sum_k w_{ik}} Y_i + \epsilon_i$$

Note on eigenvalues and eigenvectors, simple example

- Let $A = \begin{bmatrix} -1 & 2 \\ 0 & 3 \end{bmatrix}$
- If $Au = \lambda u$ (i.e. $(A - \lambda Id) * u = 0$) then u is an eigenvector associated with eigenvalue λ
- eigenvalues, solve the equation $det|A - \lambda Id| = 0$. Find $\lambda = -1$ or 3. Eigenvectors $(1, 0)$ and $(1, 2)$

How to choose λ such that $(I - \lambda W)^{-1}$ exists?

- $(I - \lambda W)$ exists iff $det(I - \lambda W) \neq 0$
- If $det(I - \lambda W) = 0$ then ($\lambda \neq 0$ and $\frac{1}{\lambda}$ is an eigenvalue of W).
- From two previous statements, $(I - \lambda W)^{-1}$ exists if ($\lambda = 0$ or $\frac{1}{\lambda}$ is not an eigenvalue of W)

Simultaneous Autoregressive **Regression** Model

$$
\begin{aligned}
Y &= X\beta + C(Y - X\beta) + \epsilon; \text{ or equivalently} \\
\text{Data} &= \text{ Linear trend } + \text{ Spatial signal } + \text{ error} \\
Y &= CY + (I - C)X\beta + \epsilon
\end{aligned}
$$

where X is a set of covariates, ϵ_i's are independent and $\epsilon_i \sim N(0, \sigma_i^2)$

Fitting SAR model

- find parameters $\hat{\beta}$, $\hat{\lambda}$ and $\hat{\sigma}_i^2$'s which maximize the likelihood.
- In spdep, $\sigma_i^2 = d_i * \sigma^2$, where the weights d_i's are provided by user and parameter σ^2 is fitted. In following results, $d_i = 1$ for all i.

Fitting SAR model

$$Y = X\beta + C(Y - X\beta) + \epsilon$$

- Fitted Values: $\hat{Y} = X\hat{\beta} + \hat{\lambda}W(Y - X\hat{\beta})$
- Residuals: $Y - \hat{Y}$
- Fitted linear trend $X\hat{\beta}$
- Fitted spatial signal $\hat{\lambda}W(Y - X\hat{\beta})$

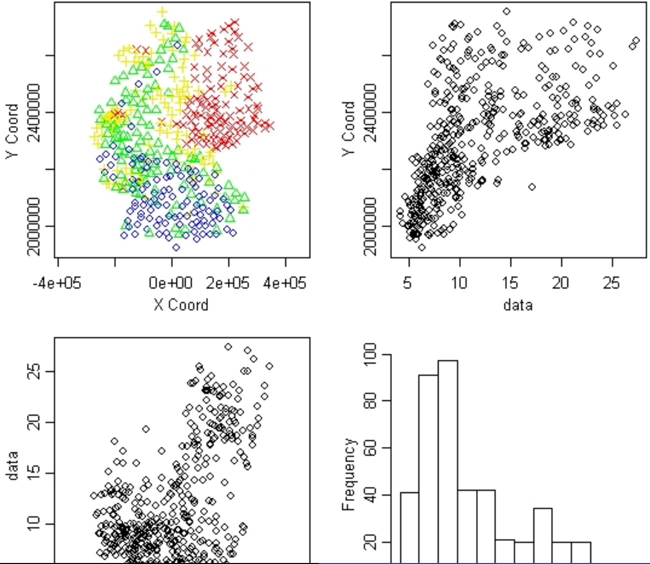
Spatial lag model

Model considered by economists (Anselin, 1988)

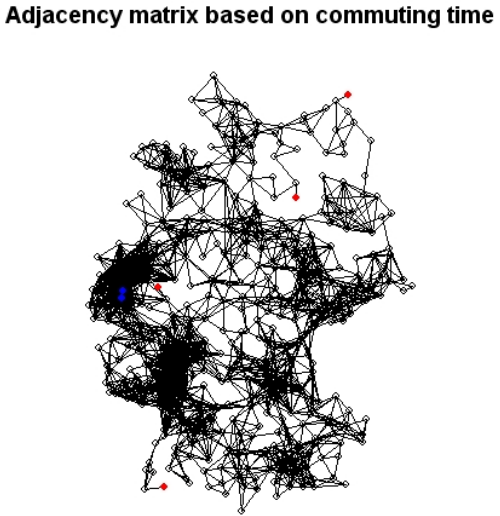
$$\begin{aligned} Y - X\beta &= C{\color{red}Y} + \epsilon; \text{ or equivalently} \\ Y &= CY + X\beta + \epsilon \\ \text{Data} &= \text{Spatial Signal } + \text{ Linear Trend } + \text{ error} \end{aligned}$$

where X is a set of covariates, ϵ_i's are independent and $\epsilon_i \sim N(0, \sigma_i^2)$

German data, EDA



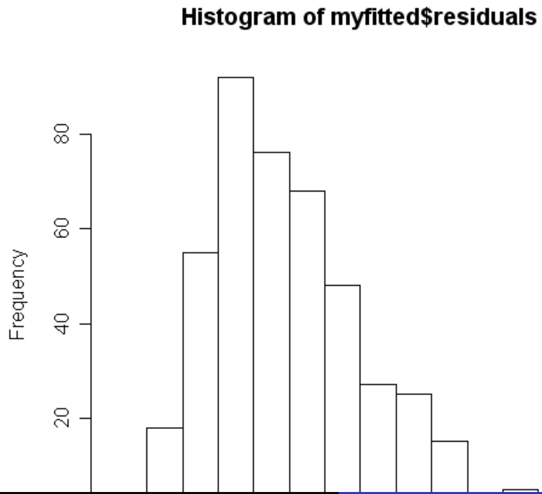
German data, graph used (commuting time less than 60min)



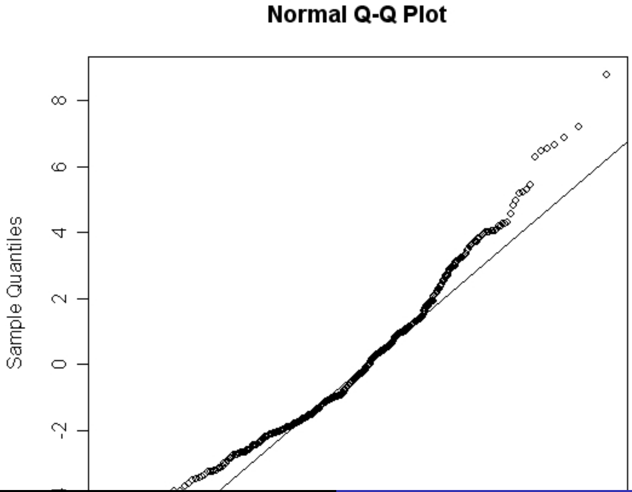
German data, output of spautolm in spdep in R with row standardized matrix

```
Call: spautolm(formula = URdata[, 4] ~ WE, data = URdata, listw = listcomm2)

Residuals:
     Min       1Q   Median       3Q      Max
-4.43755 -1.65348 -0.37015  1.20400  8.78432

Coefficients:
            Estimate Std. Error z value  Pr(>|z|)
(Intercept) 14.87249    0.96673 15.3843 < 2.2e-16
WE          -4.37039    0.79127 -5.5232 3.328e-08

Lambda: 0.86278 LR test value: 160.19 p-value: < 2.22e-16

Log likelihood: -1010.601
ML residual variance (sigma squared): 4.9872, (sigma: 2.2332)
Number of observations: 439
Number of parameters estimated: 4
AIC: 2029.2
```

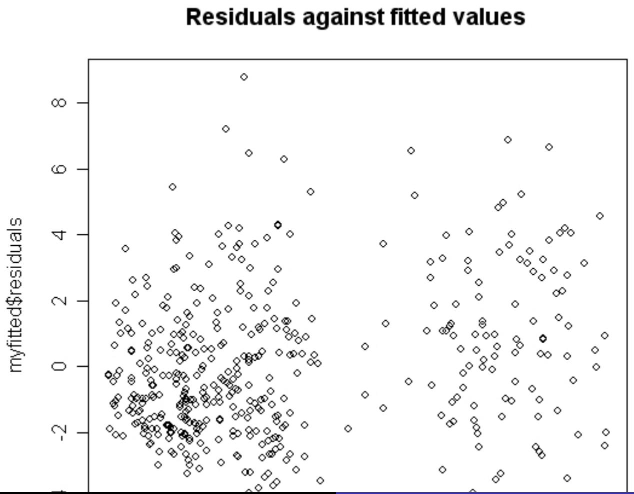
German data, inspection of residuals (SAR) (output with row standardized matrix)



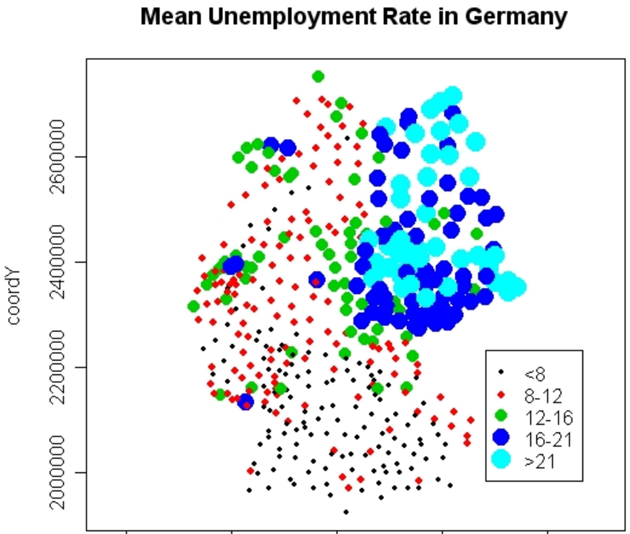
German data, inspection of residuals (SAR) (output with row standardized matrix)



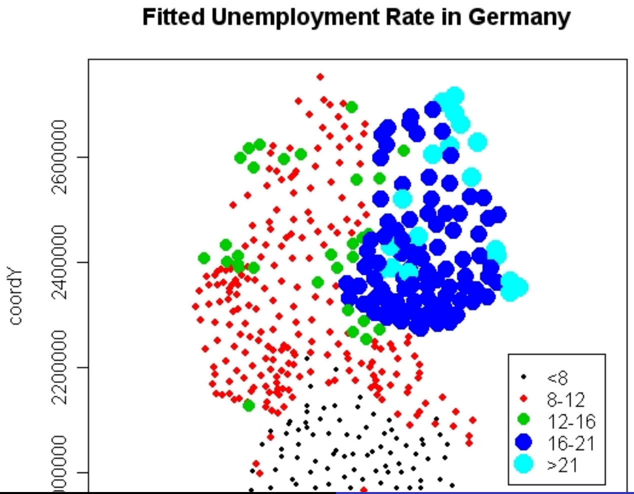
German data, residuals vs fitted values (SAR) (output with row standardized matrix)



German data, observed values



German data, Fitted values (SAR) (output with row standardized matrix)



German data, Residuals (SAR) (output with row standardized matrix)