Rental Apartment Prices in the province of Zurich

Assignment 1 for Spatial Statistics (STAT 946)

Adrian Waddell

University of Waterloo

October 9, 2008

Goal

- Overview of real estate market in Zurich
- Fit a model

$$\text{price} \sim \text{location} + \text{other covariates} + \text{error}$$

- which apartments have large residuals?
- can model be used to classify good and bad deals?
- automate process, daily update

Data Sources

Final Data: 3088 apartments for rent in province Zurich (Switzerland), collected on Friday, October 3, 2008.

street, nr, postal code, city, longitude, latitude, number of rooms, living area, apartment style, floor, price

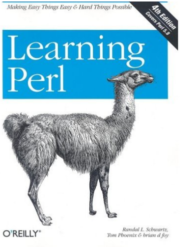
Real Estate Data	homegate.ch
Geocoding	Google Maps BETA API
GIS Data	http://www.giszh.zh.ch (CH1903)

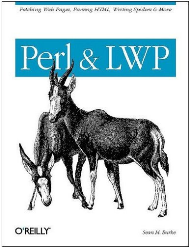
Data Collection

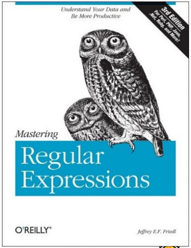
Perl Script 1: Search for all apartments in Zurich, save the html page sources for each list $\longrightarrow$ 165 $*$.txt files.

Perl Script 2: Information extraction form html sources (parsing). Lookup longitude and latitude with Google API (geocoding). (library Geo::Coder::Google).

Books on this Topic: (all O'Reilly)



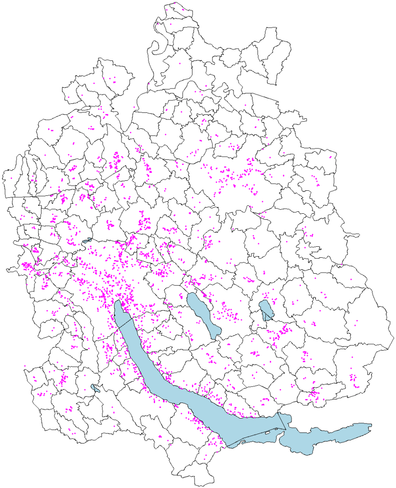




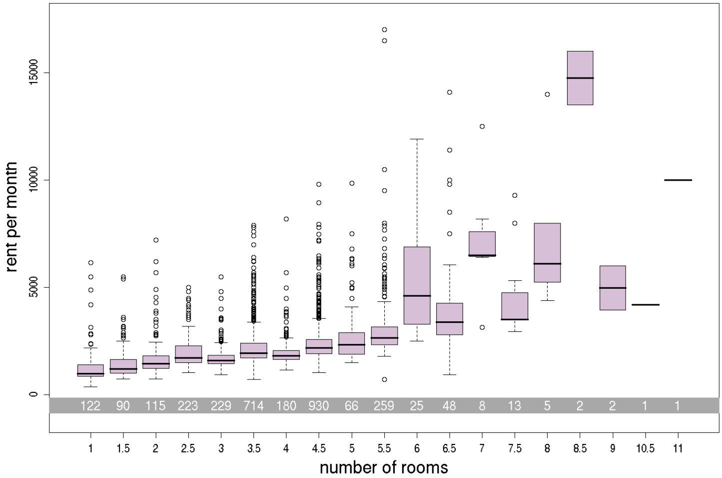
Data Processing

- All data imported into R.
- Coordinate Reference System chosen to be the "Swiss coordinate system". Transformation of housing data.
- Outliers detection (in location and price) and deletion. $3144 - 3088 = 56$ outliers.

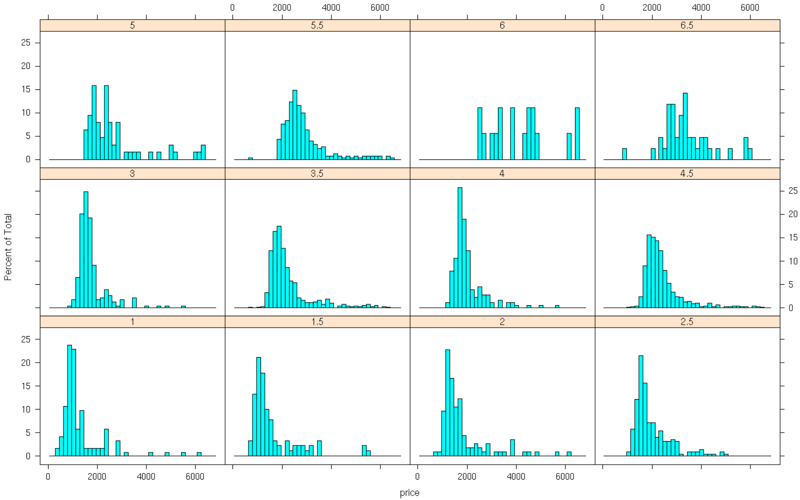
All available apartments for rent ($n = 3088$)



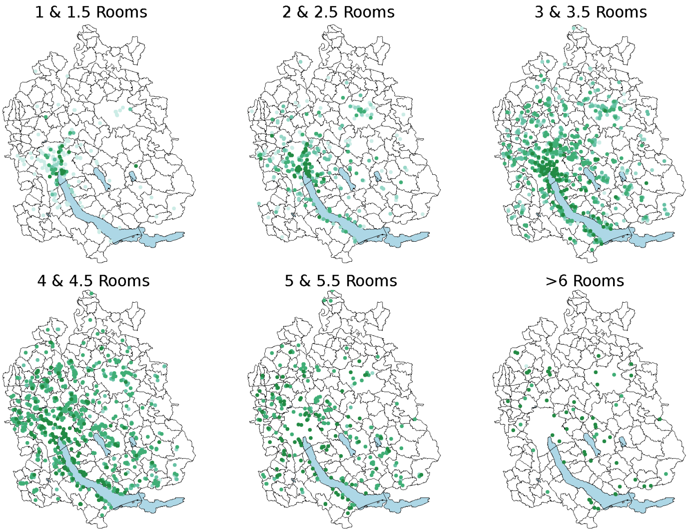
Price vs. number of rooms



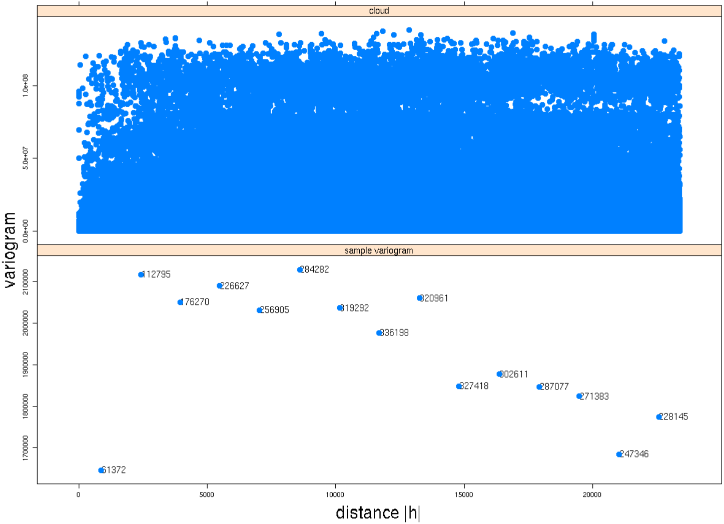
Price distribution for Nr. of Rooms $\leqslant 6.5$ and price < 6700



Price vs. number of Rooms



Is the location sufficient to explain the monthly rent?



Model

- Location is not sufficient to describe price.
- Use Model

$$\begin{aligned} \log(\mathrm{price}) &= m(\cdot) + e(s) \\ e(s) &= f(s) + \epsilon \end{aligned}$$

- **non-spatial trend:** $m(area, nrRooms, \ldots)$ is chosen to be a linear model $\rightarrow$ variable selection
- **spatial trend:** $e(s)$, model Variogram, Kriege
- **residuals:** ϵ

Variable selection: apartment style

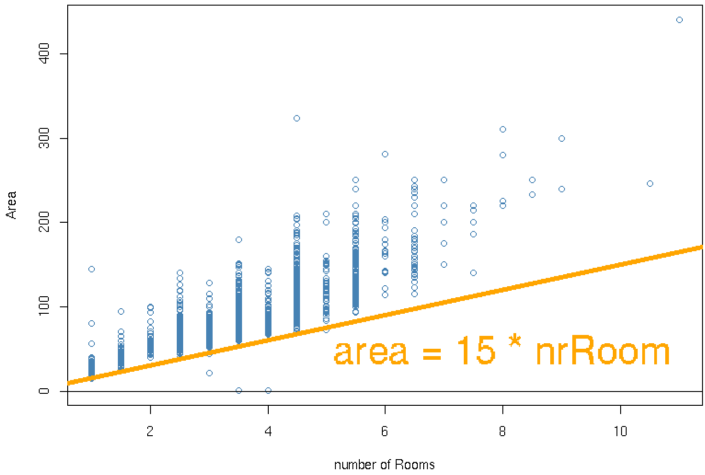
	style	Number or Rooms [1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,12)	Not Avail
*	Apartment	114	228	750	873	201	26	24
	Attic	1	0	0	0	0	0	0
*	Attic flat	5	8	27	36	17	3	0
	Bachelor flat	0	2	0	0	0	0	0
	Bifamiliar house	0	0	2	3	3	4	0
*	Duplex	1	14	40	101	51	14	2
	Farm house	0	0	1	1	1	4	0
*	Furnished flat	67	59	62	22	5	3	13
	Loft	5	1	2	2	0	0	10
*	Roof flat	4	25	55	44	15	2	2
*	Row house	1	0	1	15	16	14	1
*	Single house	0	0	1	9	11	31	0
	Single room	10	1	0	1	0	0	2
	Studio	4	0	0	0	0	0	1
	Terrace flat	0	0	2	3	4	0	0
	Terrace house	0	0	0	0	0	1	0
	Villa	0	0	0	0	1	3	0

Variable selection: apartment are

	area available	
nr Room	YES	NO
[1,2)	163	49
[2,3)	275	63
[3,4)	791	152
[4,5)	937	173
[5,6)	283	42
[6,12)	96	9
Not Avail	37	18
total	2582	506

- Only use apartments with styles marked with * ($n = 3013$)
- Only use apartments with available living area data

Variable selection summary



Model fitting

- Use $area$, $style$ and $nrRoom$ as covariates
- Omit NA's and $nrRoom > 6.5$, $area > 5 \rightarrow n = 2464$
- Fit linear model

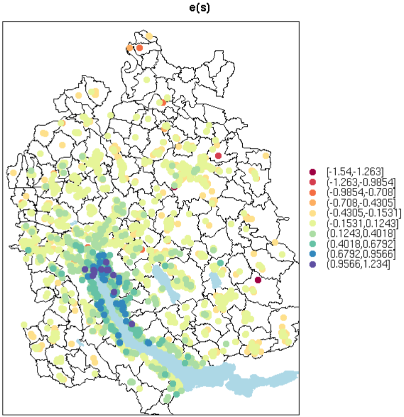
$$\log(price) = \beta_0 + \beta_1 \cdot area + \beta_2 \cdot nrRooms + \beta_3 \cdot style + e(s)$$

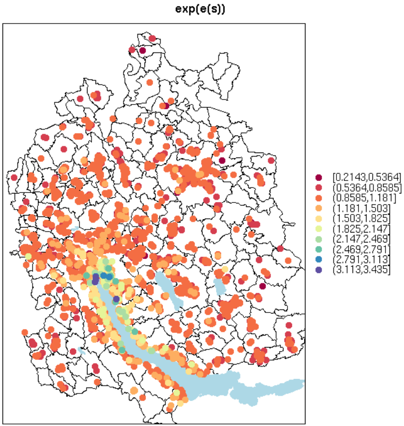
where $nrRooms$ and $style$ are factor variables.

Fitted Model

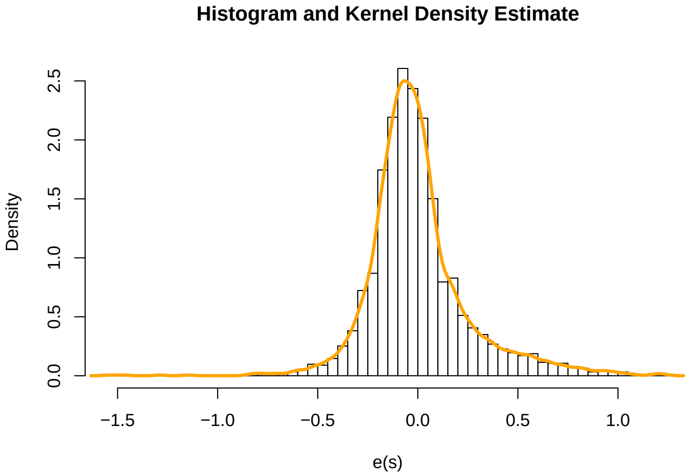
```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)          6.6575610  0.0313623 212.279  < 2e-16 ***
area                 0.0075245  0.0002692  27.951  < 2e-16 ***
nrRoom:1.5           0.1374210  0.0419900   3.273  0.00108 **
nrRoom:2             0.2065559  0.0413576   4.994 6.32e-07 ***
nrRoom:2.5           0.2818575  0.0365278   7.716 1.73e-14 ***
nrRoom:3             0.2314567  0.0372024   6.222 5.77e-10 ***
nrRoom:3.5           0.2923112  0.0353915   8.259 2.37e-16 ***
nrRoom:4             0.2188876  0.0401093   5.457 5.32e-08 ***
nrRoom:4.5           0.2421336  0.0381684   6.344 2.66e-10 ***
nrRoom:5             0.2953283  0.0511765   5.771 8.89e-09 ***
nrRoom:5.5           0.2279178  0.0450000   5.065 4.39e-07 ***
nrRoom:6             0.4685403  0.0738201   6.347 2.61e-10 ***
nrRoom:6.5           0.2776106  0.0624401   4.446 9.14e-06 ***
style:Attic flat     0.2061413  0.0288673   7.141 1.22e-12 ***
style:Duplex         0.0008961  0.0204669   0.044  0.96508
style:Furnished flat 0.5765866  0.0217763  26.478  < 2e-16 ***
style:Roof flat     -0.0006020  0.0236714  -0.025  0.97971
style:Row house     -0.1118195  0.0509342  -2.195  0.02823 *
style:Single house   0.1376427  0.0504790   2.727  0.00644 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.254 on 2445 degrees of freedom
Multiple R-squared: 0.5796, Adjusted R-squared: 0.5765
F-statistic: 187.3 on 18 and 2445 DF,  p-value: < 2.2e-16
```

Spatial trend: $e(s)$ & $\exp\{e(s)\}$

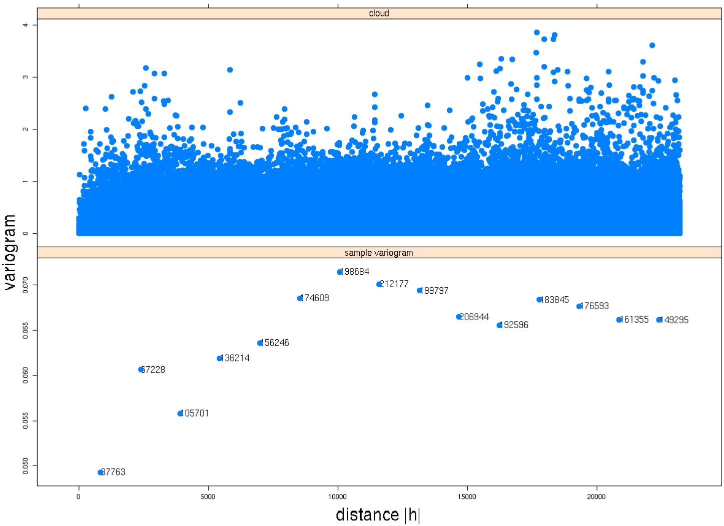




Distribution of $e(s)$



Omnidirectional Variogram (MoM) for $e(s)$

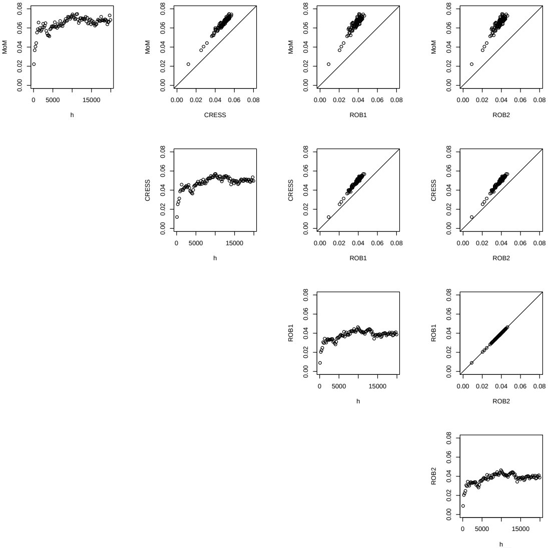


Robust Variogram estimates

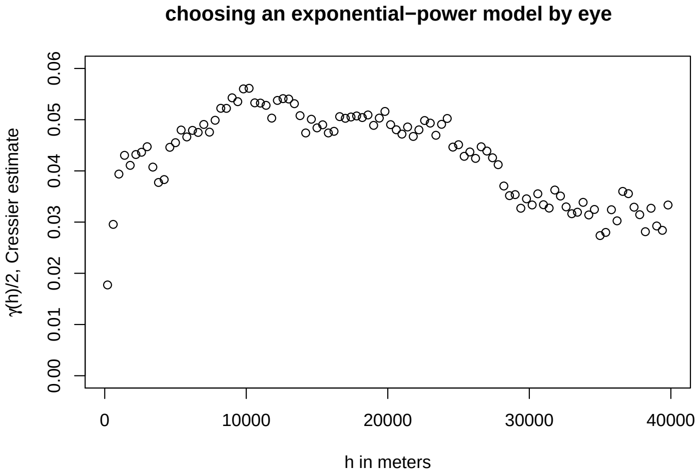
$$\begin{aligned}
MoM(h) &= \frac{1}{2} \cdot \frac{1}{|N(h)|} \sum_{(s_i,s_j)\in N(h)} \{e(s_i) - e(s_j)\}^2 \\
CRESS(h) &= \frac{1}{2} \cdot \frac{1}{0.457 + 0.494/|N(h)|} \left\{ \frac{1}{|N(h)|} \sum_{(s_i,s_j)\in N(h)} |e(s_i) - e(s_j)|^{1/2} \right\}^4 \\
ROB1(h) &= \frac{1}{2} \cdot \frac{\text{Median}[\{e(s_i) - e(s_j)\}^2 : \ (s_i, s_j) \in N(h)]}{0.457} \\
ROB2(h) &= \frac{1}{2} \cdot \frac{\text{Median}[\{e(s_i) - e(s_j)\}^{1/2} : \ (s_i, s_j) \in N(h)]^4}{0.457}
\end{aligned}$$

as defined in the course notes.

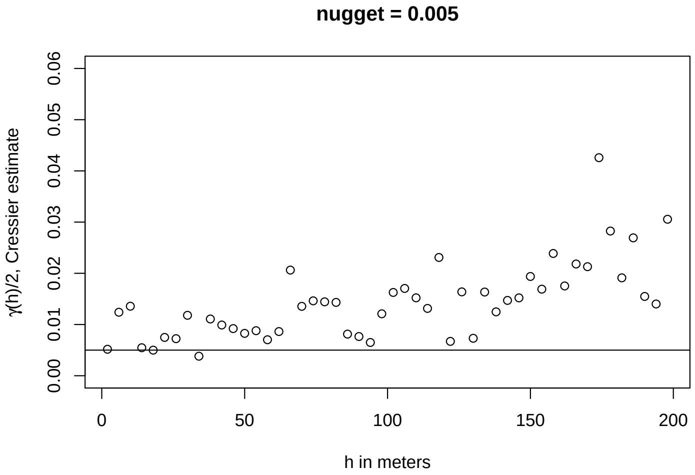
Robust Variogram estimates



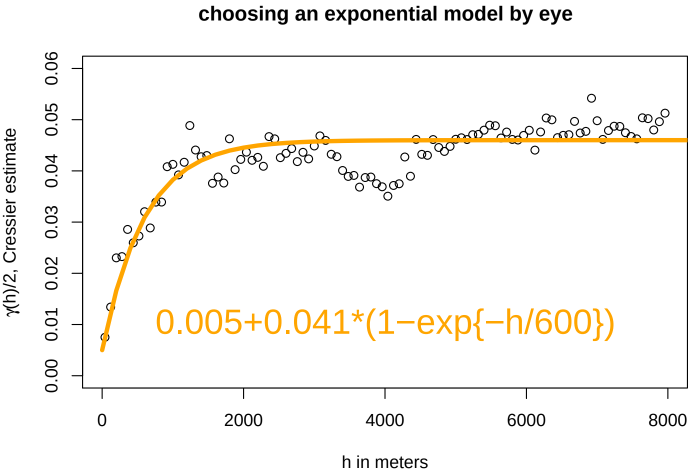
Variogram Modeling: up to h = 40km



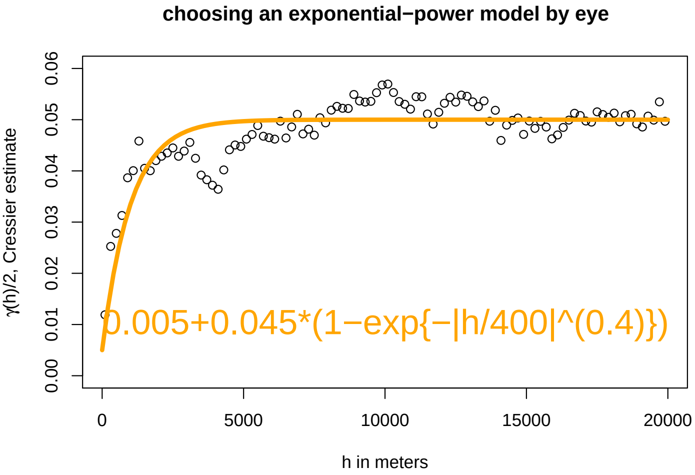
Variogram Modeling: Nugget?



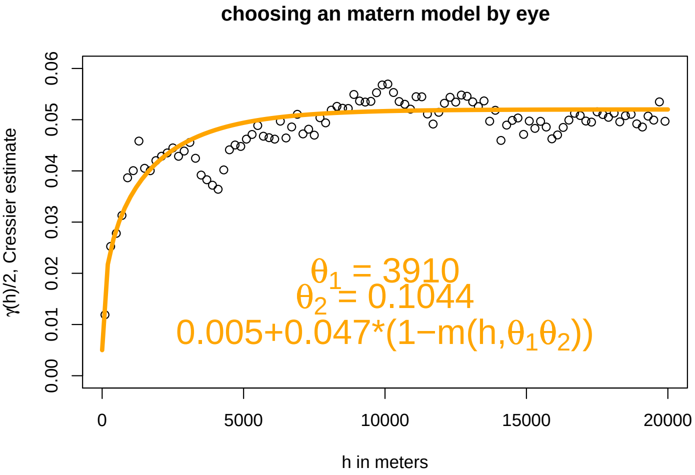
Variogram Modeling: Fitting bye eye up to h = 8km



Variogram Modeling: Fitting bye eye up to h = 20km

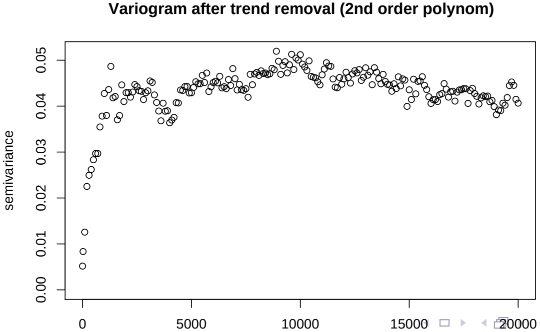


Variogram Modeling: Fitting bye eye up to h = 20km

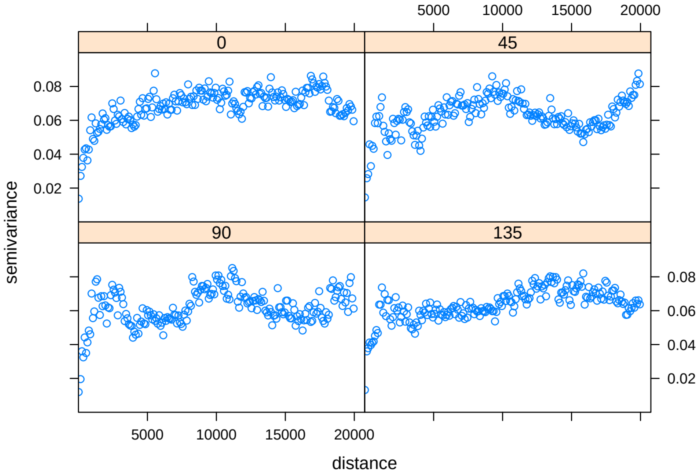


Intrinsic Stationary? Weak Stationary?

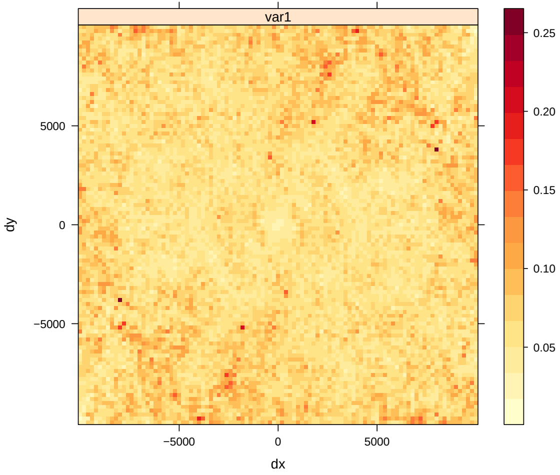
- $\gamma(h)$ flattens as h gets larger, $\mathrm{Cov}(e(s+h), e(s))$ goes to 0 as h goes to a large distance
- If data is intrinsic then it is also weak stationary.
- However looks like the mean is not constant for all locations s.
- Data may be weak stationary
- More investigation has to be done.



Directional Variograms



Directional Variograms: Variomap



Fit of empirical variogram with OLS

- Model chosen: Matern, nugget $= 0.005$ fixed, θ_2 variabel, initial values : $\sigma^2 = 0.05$ and $\phi = 2000$

- OLS Fit

  $\gamma_{ols}(h) = 0.005 + 0.0442 \cdot (1 - \mathrm{matern}(h, \theta_1 = 440.656, \theta_2 = 1))$

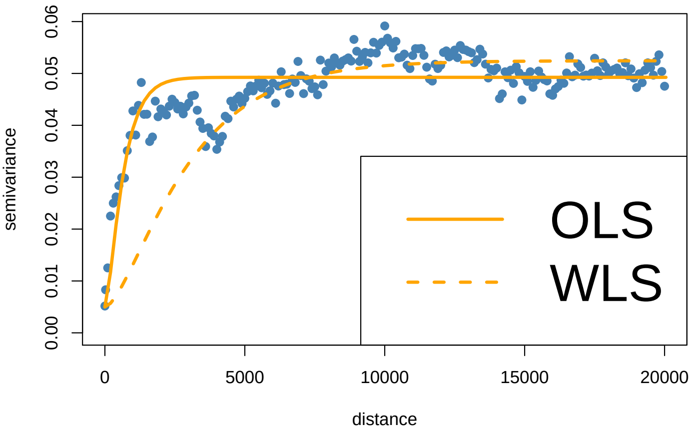
- WLS Fit

  $\gamma_{WLS}(h) = 0.005 + 0.047 \cdot (1 - \mathrm{matern}(h, \theta_1 = 1999, \theta_2 = 1))$

- Sum of Squares: 0.00344686 and 54.92099

- Practical Range: 1761.974 and 7997.04

Fit of empirical variogram with OLS and WLS



ML and REML

- Data set too large to run ML and REML
- Sampling doesn't yield good results
- cutoff can't be specified

Discussion

Results:

- Data may be weakly stationary
- Data is likely to be isotopic
- Data may be homogeneous
- Variogram Model fit by eye, Matern looks best
- Range of 1.5km-5km makes sense (size of a township)

Todo:

- In more detail analysis of trend.
- Maybe more complex non-spatial model (with postal code as covariate)

End

THANK YOU