

# Praktikum z Ekonometrie

Prostorové modely na trhu Pražských bytů

Petr Hrobař

Seminární práce

Vysoká škola ekonomická v Praze

## 1 Úvod

V této seminární práci budeme aplikovat prostorové modely na trhu pražských bytů. Uvažujeme, že cena není pouze funkcí vlastních charakteristik každé nemovitosti, ale také funkcí lokality. V návaznosti na *Toblerovo pravidlo geografie*: o podobnosti sousedních jednotek, modelujeme cenu nejen jako funkci vlastních charakteristik, ale také charakteristik sousedních jednotek.

Poznatky naší studii jsou následující: Nejdříve na základě parametru prostorové autokorelace (statisticky signifikantní na 1% hladině) potvrzujeme prostorovou závislost, tj, že ceny pražských nemovitostí nejsou náhodně rozmištěny v prostoru, ale tvoří určité *prostorové clustery*, kde ceny a charakteristiky nemovitostí jsou si více podobné, než mimo ně.

Dále jakmile bereme v potaz prostorovou závislost, jsme schopni ceny nemovitostí odhadovat přesněji. Nakonec dle distribuce reziduů identifikováváme *honosné clustery*, kde cena nemovitostí může být i dvojnásobně větší než její predikce, čistě z důvodů lokality.

Formálně si stanovíme následující výzkumné hypotézy:

- H1: *Z důvodu prostorové autokorelace disponují prostorové modely lepší predikční schopnosti.*
- H2: *Historické centrum Prahy představuje hlavní „honosný“ cluster.*
- H3: *Novostavba výrazně zvýší cenu*

## 2 Dataset

Kompletní dataset získaný pro analýzu pochází z internetových stránek realitní kanceláře [www.sreality.cz](http://www.sreality.cz). Vlastní data byly extrahovaná z webu dne 10. března 2020 využitím techniky *webscrapingu* v programovacím jazyku PYTHON. V případě zájmu je dataset a výpočetní kódy k dispozici u autora seminární práce.

Z webu bylo extrahováno více než 5 000 záznamů o více než 100 proměnných. Pro účely naší analýzy bylo nutné proměnné, které nejsou pro naši analýzu relevantní, vyjmout. Dataset tedy obsahuje pro každou nemovitost následující proměnné: *Cena (price)*, *Počet pokojů (Rooms)*, *Metry čtvereční (Meters)*, *Mezonový byt (Mezone)*, *Kuchyňský koutek (KK)*, *Panelový typ (Panel)*, *Balkón/Terasa (balcony or terrace)*, *Novostavba (Novostavba)*, *Souřadnice*.

Dále bylo nezbytné odstranit všechna pozorování, která obsahovala neúplné záznamy ve výše vypsaných proměnných. Celkem datový soubor obsahuje 2 984 kompletních záznamů.

## 2.1 Základní charakteristiky

V rychlosti nahlédneme na popisné statistiky našich proměnných. Jednotlivé relevantní statistiky jsou popsány v tabulce ?? níže:

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
price	2,984	8,495,653.000	5,696,973.000	80,000	4,990,000	9,990,000	79,000,000
Rooms	2,984	2.804	1.128	1	2	4	6
Meters	2,984	79.634	40.968	15	52	96	435
Mezone	2,984	0.027	0.163	0	0	0	1
KK	2,984	0.792	0.406	0	1	1	1
panel	2,984	0.132	0.338	0	0	0	1
balcony_or_terrass	2,984	0.514	0.500	0	0	1	1
metro	2,984	0.979	0.143	0	1	1	1
novostavba	2,984	0.291	0.454	0	0	1	1

Tabulka 1: Popisné statistiky proměnných

Dále také testujeme jednotlivé vysvětlující proměnné na výraznou korelaci mezi sebou, abychom dodrželi předpoklad na absenci *dokonalé multikolinearity*. Je zřejmé, že lze předpokládat výraznou korelaci mezi proměnnými *Meters* a *Rooms*. Uvažujeme ale, že proměnné nejsou *dokonale korelované*<sup>1</sup> a do modelů budeme využívat obě.

Dále lze nahlédnout na histogramy proměnných na obrázku ?? v sekci ???. Distribuce proměnných v prostoru je možné sledovat na obrázku ?? v sekci ??.

## 3 Metodologie

### 3.1 Využitý model

K ověření našich stanovených hypotéz sestavíme model, který bude mít následující podobu:

$$\log(price) = \beta_0 + \beta_1 rooms + \beta_2 \log(meters) + \beta_3 mezon + \beta_4 kk + \beta_5 panel + \beta_6 terasa/balkon + \beta_7 novostavba + \varepsilon. \quad (1)$$

Z histogramů proměnných ?? lze usuzovat, že proměnné *Price* a *Meters* pocházejí z *log-normálního rozdělení* tj., že logaritmus proměnné má rozdělení *normální*. Obě proměnné tedy logaritmujeme. Tato forma modelů také poslouží k potlačení případně *heteroskedasticity* a lepšímu popisu vztahů v proměnných.

Model ?? budeme odhadovat několika metodami: *Metody nejmenších čtverců (OLS)*, *Kvantilová regrese*, *Spatial lag model*, *Spatial error model*.

<sup>1</sup>Tato korelace nabývá hodnoty 0.726.

## 3.2 Základní modely

### 3.2.1 OLS model

V rychlosti připomeňme regresi s využitím metody nejmenších čtverců. Minimalizujeme kvadráty reziduí dle vzorce:

$$RSS = \sum_{i=1}^n \left( y_i - \beta_0 - \sum_{j=1}^p \beta_j x_{ij} \right)^2,$$

Kde řešení je k dispozici v analytickém tvaru:

$$\hat{\beta} = (X^T X)^{-1} X^T y.$$

U čtenářů očekáváme pokročilou znalost metody OLS a nebude zde opakovat podrobné vlastnosti a předpoklady odhadu. Pro rychlou rekapitulaci lze odkázat na literaturu např. (? , ?).

Dále abychom vzali částečně v potaz prostorovou závislost využijeme proměnných *souřadnic*. Obě proměnné není vhodné přímo vložit do modelu, ale lze využít Clusterovacího algoritmu a následně do modelu vložit identifikátor Clusteru. Pro naší studii využijeme *k-means* algoritmu (? , ?). Pro ilustraci fungování Clusterování souřadnic lze nahlédnout na obrázek ???. Do modelu bude následně vložena Dummy proměnná každého Clusteru. Pokud proměnné identifikátoru *Clusteru* budou signifikantní, lze toho využít jako určitého indikátoru *prostorového shluku*.

### 3.2.2 Kvantilová regrese

Metoda kvantilové regrese je metoda ve které se minimalizuje následující tvar:

každé  $i$ -té reziduum má podobu:

$$e_i = (y_i - x_i \hat{\beta}_q),$$

pro každý konkrétní  $q$ -tý kvantil odhadujeme parametry  $\hat{\beta}_q$  takové, aby se minimalizoval výraz:

$$\min : Q_n(\hat{\beta}_q) = \sum_{i: e_i \geq 0}^n q|y_i - x_i \hat{\beta}_q| + \sum_{i: e_i \leq 0}^n (1-q)|y_i - x_i \hat{\beta}_q|.$$

Řešení výše není k dispozici v analytickém tvaru a úloha se řeší iteračně jako úloha lineárního programování (? , ?).

Je dobře známo, že metoda je robustní vůči odlehlym pozorováním. Jelikož náš datový soubor také obsahuje velké množství pozorování ( $n = 2\,984$ ) lze očekávat vhodné asymptotické vlastnosti. Abychom také ověřili stabilitu koeficientů pro různé hodnoty kvantilů. Provedeme regresi pro jednotlivé decily datového souboru.

### 3.3 Prostorové modely

Než začneme aplikovat prostorové modely, je nezbytné ověřit prostorovou autokorelaci. Standartní testy pro přítomnost prostorové autokorelace ověřují souvislosti mezi polohou pozorování hodnotami proměnných.

Výhodou prostorových modelů je možnost kontroly prostorové závislosti v data generujícím procesu (DGP). Nicméně, aby bylo možné statistické inference a odhadu koeficientů byly co nejpřesnější, je nezbytné korektní specifikování prostorové závislosti.

#### 3.3.1 Spatil lag model

*Spatil lag* model bere v potaz prostorovou závislost mezi vysvětlujícími proměnnými. Formálně lze model přepsal do podoby:

$$y = \rho W y + X\beta + \varepsilon, \quad (2)$$

Kde parametr  $\rho$  představuje parametr prostorové autokorelace. Z předpisu je patrné, že pokud:  $\rho = 0$ , pracujeme s běžným OLS modelem. Řešení ?? není k dispozici v analytickém tvaru a jako odhadová metoda parametrů modelu se užívá metoda *maximální věrohodnosti*. Pro větší detaily lze nahlédnout např. (?), (?), (?), (?).

Dále parametr  $W$  představujeme *matici sousednosti*. Jedná se o matici o rozměru  $n \times n$ , která udává, která pozorování jsou definována jako sousední a která nikoliv. Matice sousednosti  $W$  představuje hlavní prvek prostorové ekonometrie (?), (?). Jelikož dochází k nadefinování matice před samotným odhadem modelu je nezbytná její korektní specifikace.

Mezi nejčastější metody nadefinování relace sousednosti lze zmínit metodu: *K-nejbližších sousedů*, *Maximální vzdálenost* a v případě pozorování v rámci regionů nebo zemí, kdy pracujeme s *Polygony* také metody *Rook* a *Queen* sousednosti. Tím se dostáváme zpět k problematice testování prostorové autokorelace, korektní nadefinování sousednosti je nezbytné, neboť matice  $W$  vstupuje také do testů prostorové autokorelace.

### 3.3.2 Spatial Error model

Model typu *Spatial error* bere v potaz prostorovu závislost náhodně složky modelu a né proměnou vysvětlující. Formálně lze model zapsat do následující podoby:

$$y = X\beta + u, \quad (3)$$

$$u = \lambda W u + \varepsilon,$$

U modelu můžeme opět testovat zda parametr  $\lambda$  nabývá hodnoty různé od 0. Pokud by parametr nabýval hodnoty 0, výraz zmizí a opět pracujeme se základním OLS modelem. Stejně jako u *Spatial Error* modelu není řešení modelu ?? k dispozici v analytickém tvaru a opět užíváme metody *maximální věrohodnosti*, kde maximalizujeme *věrohodnostní funkci*. Pro rychlou revizi metody maximální věrohodnosti lze nahlédnout do (? , ?). Pro Detailnější popis prostorové spatial error modelu lze nalézt v (? , ?), (? , ?). Výhodu modelu je, že koeficienty lze interpretovat, jako mezní efekty.

### 3.3.3 Prostorová autokorelace a relace sousednosti

Mezi nejužívanější test prostorové autokorelace patří *Moranův I test* (? , ?), který lze zapsat následovně:

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \times \frac{\varepsilon' W \varepsilon}{\varepsilon' \varepsilon}, \quad (4)$$

kde  $n$  je počet pozorování,  $w_{ij}$  představuje prvky matice  $W$ , dále  $\varepsilon$  představuje vektor reziduí z OLS modelu.

V případně nulové hypotézy o prostorové nezávislosti nabývá statistika:

$$E(I_0) = \frac{-1}{n-1}, \quad (5)$$

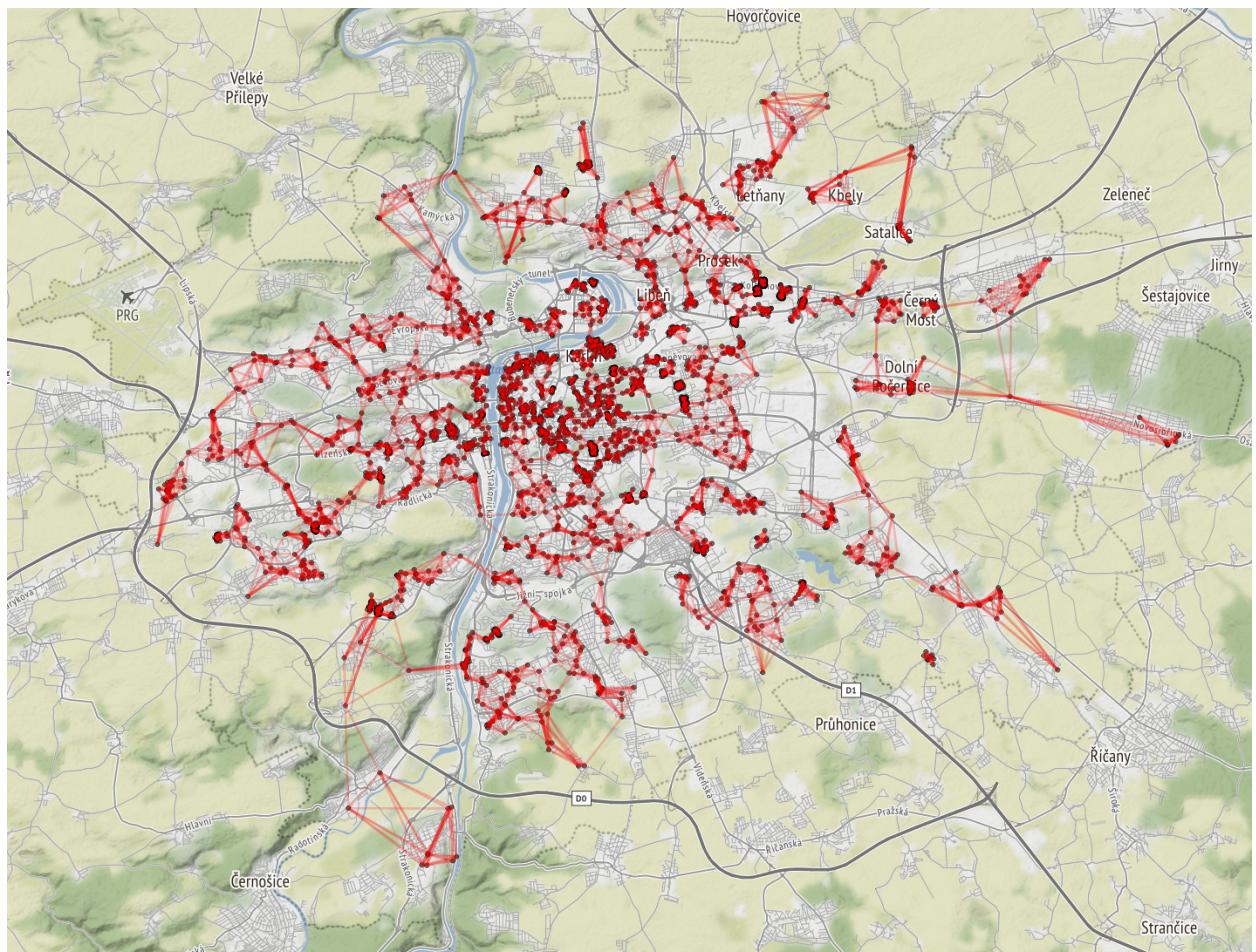
V simulačních experimentech se *Moranův test* prokazatelně odlišoval s ostatními testy v kontextu přesnosti (? , ?). Relace sousednosti můžeme také na definovat prvotně na základě úvahy.

Jelikož modelujeme ceny nemovitostí a předpokládáme, že ceny se vzájemně ovlivňují lokálně, uvažujeme, že sousední nemovitosti o stejných charakteristikách budou mít podobnou cenu, která bude také zahrnovat faktor lokality.

Budeme tedy vytvářet relaci sousednosti dle *Počtu nejbližších sousedů*. Prvotně volíme hodnotu  $k = 7$ .

Matice sousednosti může být vizualizována jako relace sousednosti pro každou pozorovanou jednotku v prostoru. Nahlédneme na relaci sousednosti na obrázku ??:

### Prostorová síť metodou 7–NN



Obrázek 1: Relace sousednosti (*7 nejbližších sousedů*)

Následně pro zvolenou relaci sousednosti provedeme *Moranův Test*, formální test a jeho výstup je možné sledovat v tabulce ??:

<i>I</i> -statistika	Rozptyl	p-value
0.413	0.0001	0.000

Tabulka 2: Moranův Test 7-NN

je patrné, že v data generujícím procesu se vyskytuje statisticky signifikantní prostorová autokorelace a volíme metodiku prostorových modelů.

O prostorové závislosti lze také uvažovat dle OLS modelů s *Kmeans* proměnnou v tabulce ??, kde proměnné Clusterů jsou statisticky signifikantní.

## 4 Empirická část

Výše v rychlosti popisujeme všechny modely a modifikace, které budeme v naší studii využívat, odhadu dílčích koeficientů všech modelů je možné sledovat v tabulce ??:

	Vysvětlovaná proměnná: <i>log(price)</i>					
	OLS		<i>quantile</i>		Prostоровé	
	<i>k-means</i>		<i>K-means</i>	<i>Spatial Lag</i>	<i>Spatial Error</i>	
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Rooms</i>	0.061*** (0.010)	0.082*** (0.009)	0.053*** (0.009)	0.089*** (0.008)	0.081*** (0.007)	0.082*** (0.007)
<i>log(Meters)</i>	0.808*** (0.021)	0.731*** (0.019)	0.818*** (0.019)	0.692*** (0.018)	0.667*** (0.017)	0.685*** (0.016)
<i>Mezone</i>	-0.003 (0.031)	-0.032 (0.027)	0.0001 (0.022)	-0.041 (0.027)	-0.048* (0.025)	-0.067** (0.023)
<i>KK</i>	0.117*** (0.016)	0.162*** (0.014)	0.095*** (0.017)	0.157*** (0.013)	0.160*** (0.012)	0.176*** (0.012)
<i>Panel</i>	-0.324*** (0.016)	-0.220*** (0.015)	-0.305*** (0.012)	-0.232*** (0.013)	-0.163*** (0.013)	-0.119*** (0.015)
<i>Balcony_or_terrase</i>	-0.007 (0.011)	0.057*** (0.010)	-0.00004 (0.010)	0.054*** (0.009)	-0.030*** (0.008)	0.072*** (0.008)
<i>Novostavba</i>	-0.011 (0.011)	0.021** (0.010)	0.018* (0.011)	0.031*** (0.009)	0.047*** (0.009)	0.073*** (0.010)
<i>Kmeans<sub>dummy2</sub></i>		-0.015 (0.020)		0.015 (0.015)		
<i>Kmeans<sub>dummy3</sub></i>		0.068*** (0.018)		0.088*** (0.015)		
<i>Kmeans<sub>dummy4</sub></i>		-0.047** (0.018)		-0.006 (0.015)		
<i>Kmeans<sub>dummy5</sub></i>		0.298*** (0.019)		0.294*** (0.017)		
<i>Intercept</i>	12.145*** (0.065)	12.246*** (0.060)	12.111*** (0.056)	12.371*** (0.055)	4.554*** (0.197)	12.371*** (0.055)
$\rho$					0.508*** (0.0129)	
$\lambda$						0.826*** (0.015)
<i>n</i>	2,984	2,984	2,984	2,984	2,984	2,984
$R^2_{pse.}$	0.748	0.797	0.748	0.794	0.825	0.857

Note:

\*p&lt;0.1; \*\*p&lt;0.05; \*\*\*p&lt;0.01

Tabulka 3: Odhadu všech modelů

Z tabulky modelů ?? vidíme několik zajímavých faktů. Nejdříve, pokud porovnáme OLS model a OLS *kmeans* model, dle metriky  $R_{pse}^2$ . (viz níže) lze vidíme, že zahrnutí Clusteru (obrázek ??) do modelu, výrazně zlepšuje zachycení DGP.

Zajímavé je, že můžeme identifikovat Cluster (??, žlutý), jehož proměnná ( $Kmeans_{dummy5}$ ) nabývá hodnoty 0.298 (statisticky signifikantní), tj. že v určité část Prahy idendifikujeme Cluster, ve kterém nemovitosti mají v průměru *ceteris-paribus* o 29.8 % procent vyšší cenu, čistě z důvodu lokality. Stejná hodnota je i v modelu na *Kvantilové regresy s dummy Clustery*. Tím tedy potvrzujeme, že cena je výrazně ovlivněna nejen vlastními charakteristikami, ale také faktorem lokality.

Dle *Moranova testu* (tabulka ??) potvrzujeme přítomnost prostorové autokorelace a pří pohled na parametry prostorové autokorelace u *Spatial Lag* a *Spatial Error* modelu vidíme, že oba parametry jsou statisticky signifikantní a také jejich metrika  $R_{pse}^2$  vychází výrazně vyšší. Dále dle metrik v tabulce ?? je patrně, že oba prostorové modely zachycují DGP výrazně lépe.

Pro lepší komparaci modelů si vytvoříme přehled několika metrik každého modelu. Využijeme následující metriky: *AIC*, *log-likel* a  $R_{pse}$ . Poslední z uvažovaných metrik je *pseudo R*, které je spočteno následujícím způsobem:

$$R_{pse} = \text{corr}(y, \hat{y})^2,$$

Využití této metriky nám umožní zachytit predikční schopnosti každého modelu.

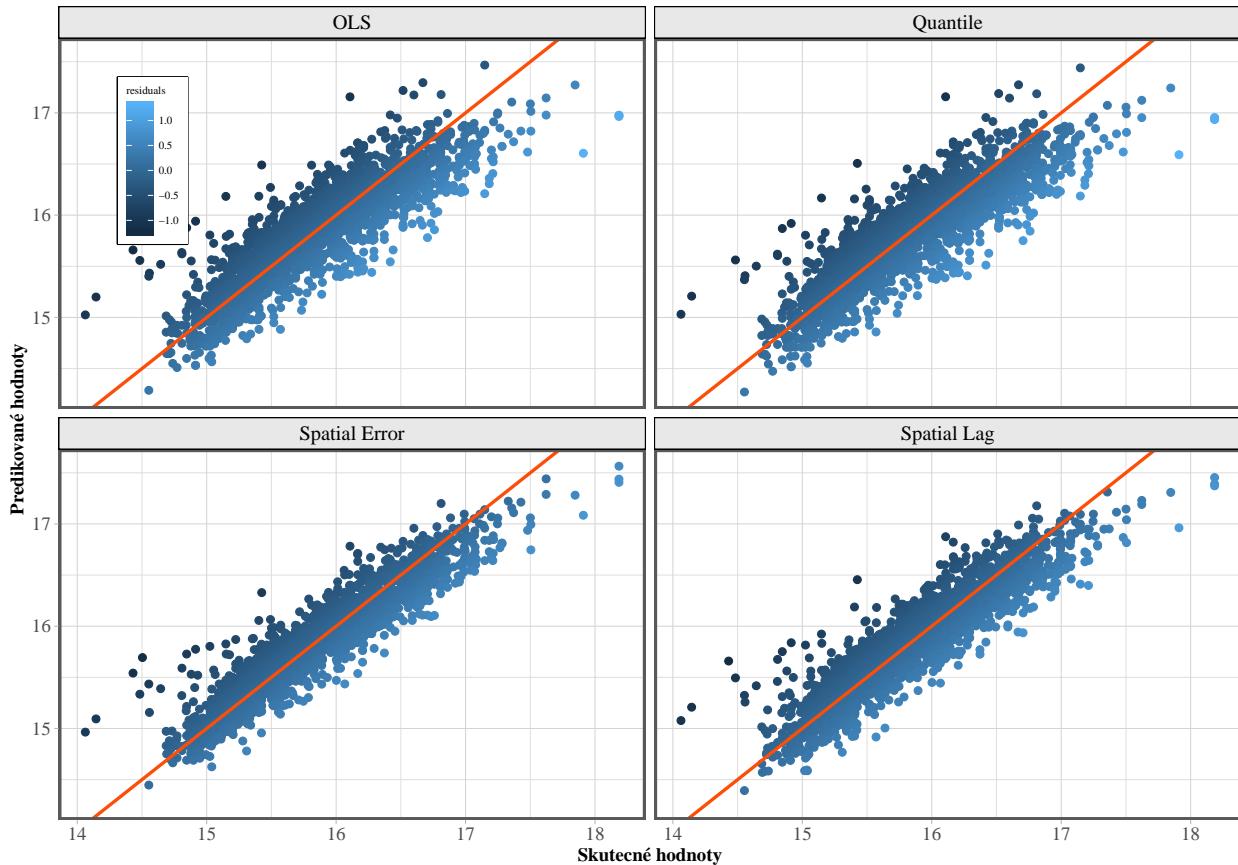
	<i>OLS</i>	<i>OLS Kmeans</i>	<i>Quantile</i>	<i>Quantile Kmeans</i>	<i>Spatial Error</i>	<i>Spatial Lag</i>
<i>AIC</i>	583.827	-15.779	29.727	-754.668	-836.191	-495.142
<i>Log-like</i>	-282.913	-20.889	-6.863	389.334	428.095	257.571
<i>pseudo R</i>	0.748	0.795	0.748	0.794	0.857	0.825
<i>n</i>	2,984	2,984	2,984	2,984	2,984	2,984

Tabulka 4: Metriky modelů

Při porovnání *OLS regrese* a *kvantilové regrese* a její citlivosti na obrázku ?? vidíme, že všechny proměnné jsou pro každý decil relativně stabilní s výjimkou proměnné *No-vostavba*.

Třebaže našim hlavním cílem práce není vytvoření *predikčního algoritmu*, ale spíše *Statistická inference* můžeme nahlédnout na porovnání skutečných a predikovaných hodnot každého modelu. Predikční přesnost modelů je možné sledovat na obrázku ??.

#### Porovnání predikce modelu



Obrázek 2: Porovnání skutečných a predikovaných hodnot

Vidíme, že hodnoty modelu *Spatial Error* jsou nejvíce sblížené s linií 45 stupňů, tedy přímkou kde skutečné a predikované hodnoty jsou zcela totožné.

Dle základě regresní tabulky ??, dále také dle metrik užitých v tabulce ?? a nakonec dle predikcí na obrázku ?? volíme *Spatial Error model* pro statistickou inferenci.

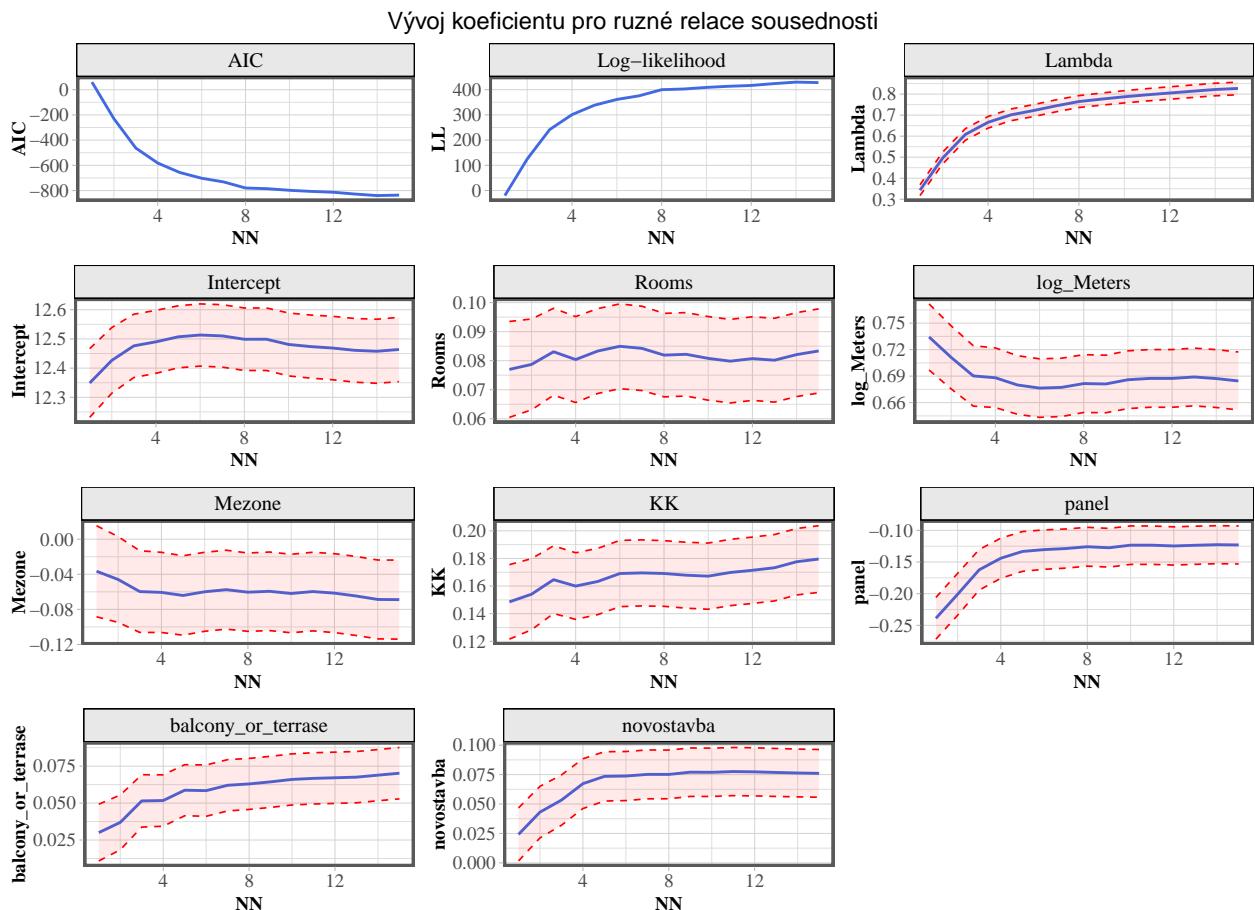
V další části práce se budeme zabývat citlivostí vůči změnám specifikace relace sousednosti a citlivosti koeficientů v závislosti na matici  $W$ .

## 5 Citlivost na zm ny relace sousednosti

V návaznosti na studii (?, ?) provedeme i my ověření stability koeficientů v závislosti na formulaci matice  $W$ . Budeme nadále uvažovat relaci sousednosti dle *K-nejbližších sousedů*, ale budeme ověřovat citlivost koeficientů pro různé hodnoty  $k$ . Hlavním důvodem proč volíme algoritmus *K-nejbližších sousedů* je skutečnost, že naše data nejsou rovnoměrně zastoupená v prostoru. pokud bychom např. volili metodiku *do maximální vzdálenosti* mohlo by se stát, že některé jednotky budou mít mnohonásobně větší počet jednotek, které jsou definované jako sousedící.

Pokud užíváme algoritmu *K-nejbližších sousedů* nebude případné zkreslení způsobené distribucí dat nakolik výrazné. Dalším důvodem je výpočetní čas. U jednotlivých prostorových modelů popisujeme, že řešení není k dispozici v analytickém tvaru a hledá se iteračně, je patrné, že pokud pracujeme s maticí  $W$ , která má v našem případě rozměr  $2984 \times 2984$  je patrné, že výpočetní náročnost na hardware počítače je značná.

Citlivost koeficientů pro různé hodnoty počtu sousedních jednotek (v našem případě od 1 do 15 jednotek) je možné sledovat na obrázku ?? níže.



Obrázek 3: Citlivost koeficientů v změny matice  $W$

V našem případě je patrné, že hodnota, která minimalizuje hodnotu informačního kritéria je  $k = 15$ , tento model také skutečně uvádíme v tabulce ?? a využíváme jej k výsledné statistické inferenci a zhodnocení stanovených hypotéz.

## 6 Distribuce reziduí v prostoru

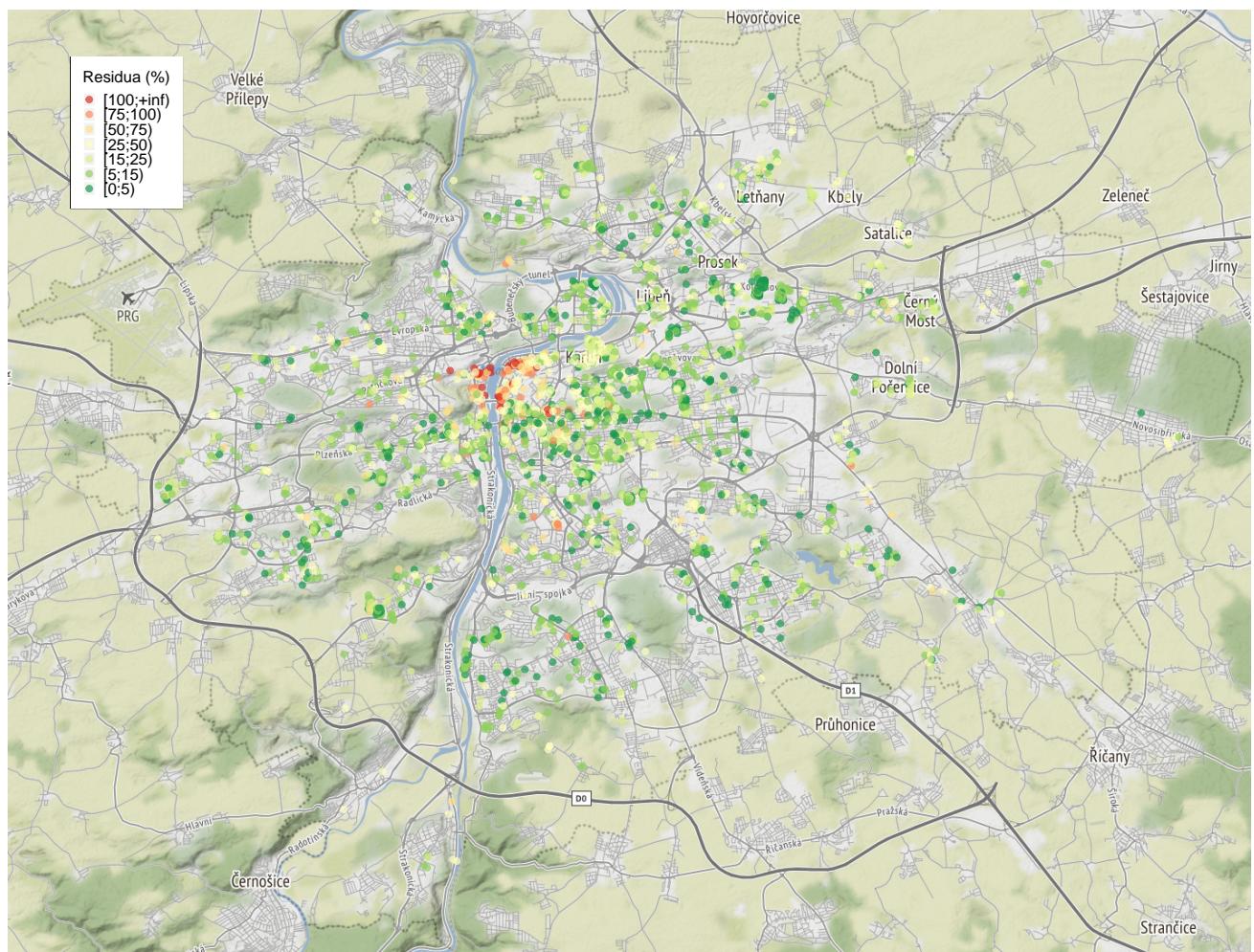
OLS model není zcela kvalitní na oceňování nemovitostí na Pražském trhu, neboť nebere v potaz prostorové závislosti. Nicméně rezidua modelu mohou resp. jejich rozmístění v prostoru může představovat zajímavý indikátor.

Nejdříve spočítáme procentuální chyby predikce z modelu jednoduché regrese (bez *Clusterů*) pro každé pozorování a následně dle intervalů hodnoty diskretizujeme.

Při pohledu na ?? níže vidíme, že nemovistosti vyskytující se v historickém centru Prahy (Staroměstské náměstí a přilehlé okolí).

Ceny nemovistostí jsou zde více jak dvojnásobné (více jak 100% rozdíl predikce) čistě z důvodů výskytů nemovistostí v historické části. Abychom tuto nevyrovnanost v reziduích odstranili, bylo by nutné každé nemovitosti v tomto centru přidat novu kontrolní proměnnou *Historické centru*, která by nabývala hodnoty 1, pro nemovitosti v *Honošném Clusteru*. Užití modelů bez prostorové závislosti nám umožňuje takovéto clustery identifikovat.

### Distribuce reziduí v Prostoru



Obrázek 4: Shluková analýza reziduí

## 7 Závěr

Na závěr naší studie se vrátíme zpět k našim stanovených hypotézám. Nejdříve jsme uvažovali, že prostorové modely z důvodu prostorové autokorelace disponují lepší predikční schopností a také lepším zachycením DGP. Tedy naší první stanovenou hypotézu lze na základě zjištění v této studii *potvrdit*.

Dále jsme uvažovali identifikaci honosných Clusterů v Pražském trhu. Hlavním ukazatelem může být distribuce reziduů z lineárního modelu, který nebere v potaz prostorovou závislost. Na obrázku ?? vidíme výrazný *Honosný Cluster* v historické části Prahy.

Takto identifikované clustery mohou mít relevantní informační hodnotu pro realitní agenty, pro obyvatelé, které o investici do nemovitosti uvažují a zejména pro majitelé nemovitostí, kteří uvažují o prodeji či pronájmu.

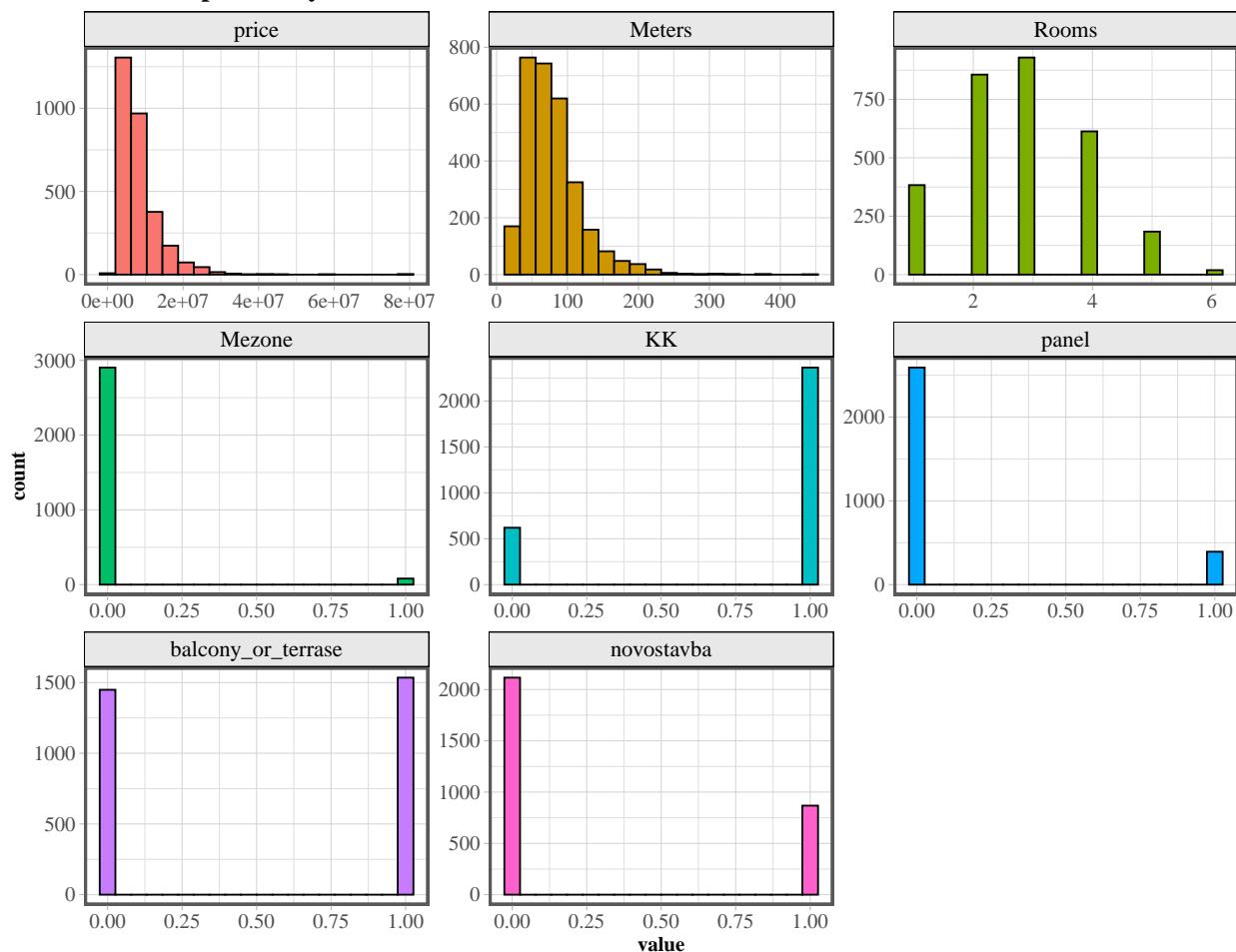
Pokud se jejich nemovitost nachází v takovémto honosném Clusteru, lze cenu nemovitosti výrazně navýšit a očekávat i tak uzavření obchodu.

K poslední stanovené hypotéze, zda novostavba relevantně zvyšuje nemovitosti využijeme *Spatial Error modelu*. Parametr *novostavby* je jak v tabulce ??, tak také na ?? vždy statisticky odlišný od hodnoty 0. Dle koeficientu u našeho modelu odhadujeme, že mezní efekt nemovitosti je roven zhruba 7.5 %, tedy opět *potvrzuje*me naši stanovenou hypotézu.

Naše Studie představuje aplikaci prostorových modelů na trhu pražských nemovitostí. Z důvodu prostorové autokorelace jsou prostorové modely pro takový typ datového souboru vhodné a jejich uplatnění v praxi stále roste. V budoucím výzkumu lze dataset rozšířit o rozdíl času a jednotlivé analýzy provádět nikoliv, jako na průřezových datech, ale na datech panelových.

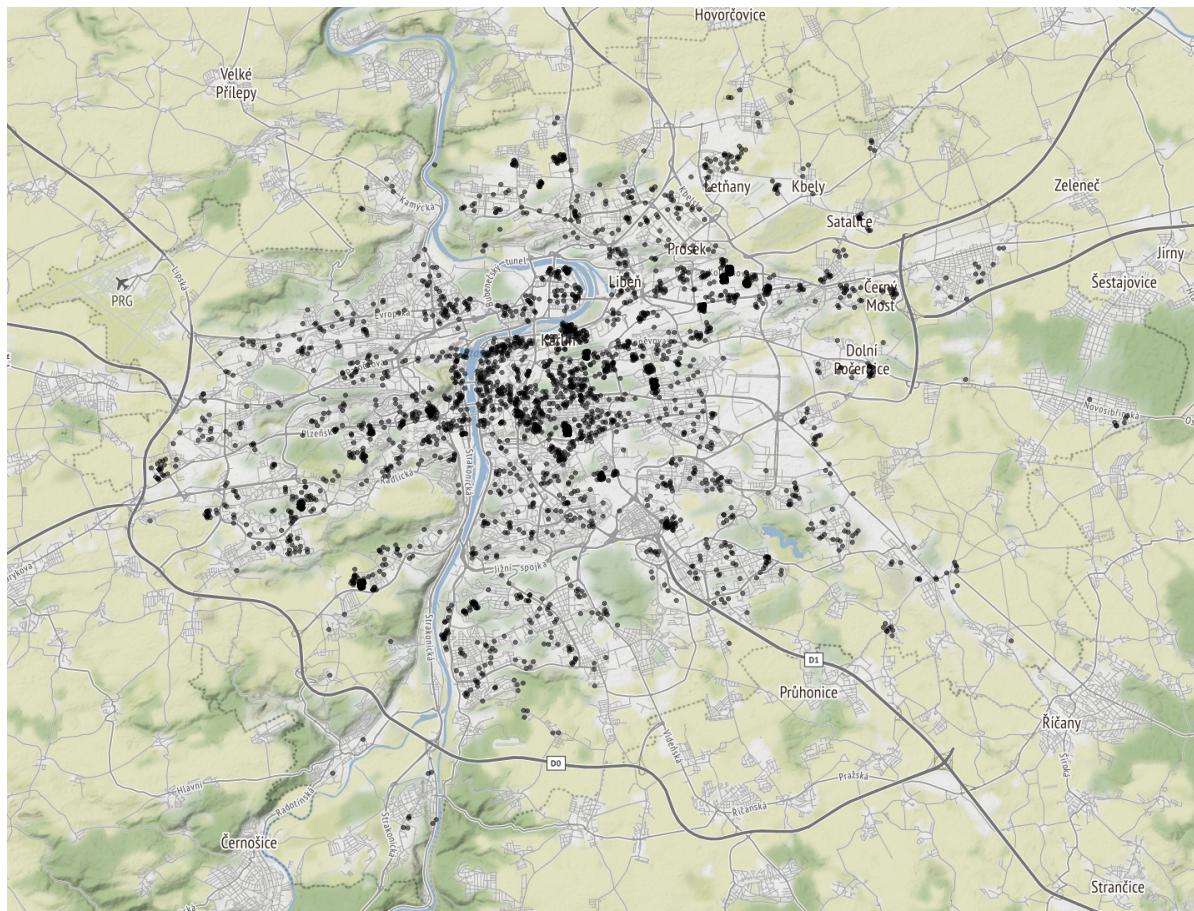
## 8 Přílohy

### Distibuce promenných



Obrázek 5: Distribuce Proměnných

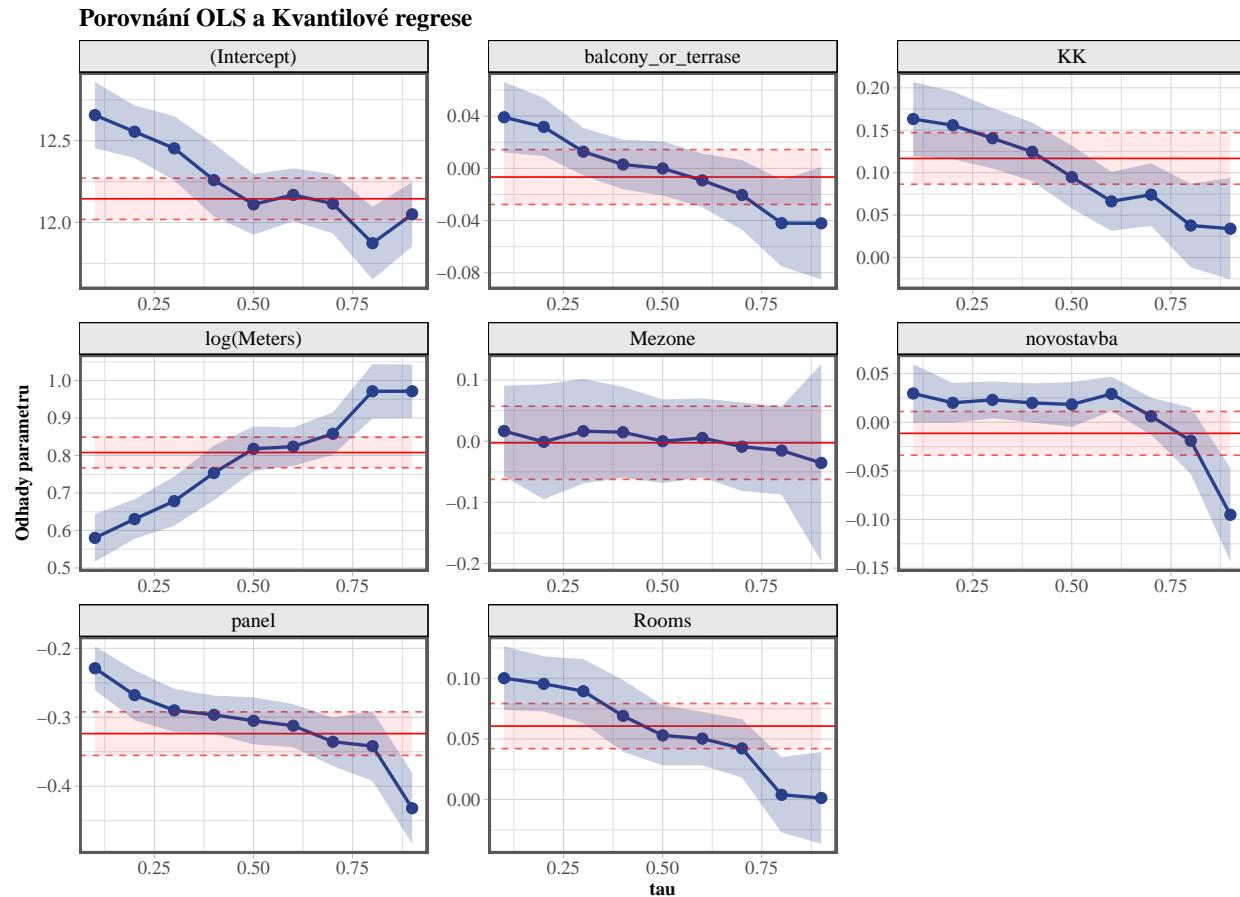
## Distribuce pozorování



Obrázek 6: Distribuce Pozorování v prostoru

"""flat\_cluster".pdf

Obrázek 7: Distribuce Pozorování a Clusterování



Obrázek 8: Citlivost Kvantilové regrese