PROSTOROVÉ REGRESNÍ MODELOVÁNÍ S PŘÍKLADY

Jiří HORÁK[1], Lucie ORLÍKOVÁ[1], Joaquín Osorio ARJONA[2], Radek SVOBODA [3]

[1] Katedra geoinformatiky, Hornicko-geologická fakulta, Vysoká škola báňská – Technická univerzita Ostrava, 17. listopadu 2172/15, 708 00 Ostrava, Česká republika
jiri.horak@vsb.cz, lucie.orlikova@vsb.cz

[2] Universidad Complutense de Madrid, Department of Geography, Madrid, Španělsko
joaquoso@ucm.es

[3] Katedra informatiky, Fakulta elektrotechniky a informatiky, Vysoká škola báňská – Technická univerzita Ostrava, 17. listopadu 2172/15, 708 00 Ostrava, Česká republika
radek.svoboda@vsb.cz

doi: https://doi.org/10.31490/9788024843988-10

Abstrakt

Prostorové regresní modelování představuje jednu z možností jak využít prostorových vazeb ve vícerozměrné statistice a rozšíření klasických regresních modelů. Cílem příspěvku je představit využití prostorového regresního modelování na 2 příkladech. Analýza nezaměstnanosti v Česku byla provedena s využitím vybraných faktorů na základě testování prostorového autoregresního modelování a geografické vážené regrese. Druhý příklad je zaměřena na vyhodnocení tweetů k dopravě ve Španělsku s využitím geografické vážené regrese.

Abstract

Spatial regression modelling with examples: Spatial regression modelling represents one of the popular possibilities how to integrate spatial relationships into multidimensional statistics and to an extension of the classic linear regression models. The aim of the paper is to demonstrate the utilization of spatial regression modelling in two case studies. The unemployment analysis in the Czech Republic employed spatial autoregressive models and a geographical weighted regression. The second case study is focused on evaluation of the spatial distribution of Spain tweets commenting public transport using a geographical weighted regression.

Klíčová slova: prostorová regrese; SAR; geografická vážená regrese; nezaměstnanost; veřejná doprava; Twitter

Keywords: spatial regression; SAR; Geographically Weighted Regression; unemployment; public transport; Twitter

1. ÚVOD

Prostorové regresní modelování představuje slibné rozšíření klasických regresních modelů s využitím prostorové struktury dat. Snaží se tím adresovat jeden z hlavních nedostatků klasických regresních modelů pracujících s prostorovými daty, tedy jejich nedostatek nezávislosti mezi pozorováními v blízkých místech, což způsobuje porušení jednoho ze základních principů řešení regresních rovnic.

Současná nabídka programových implementací prostorového regresního modelování je poměrně široká a stále více uživatelů ji využívá pro své analýzy. Ne všichni si ale uvědomují vhodnost některých postupů a jejich výsledky tím mohou být nepříjemně poznamenány.

Cílem tohoto příspěvku je poskytnout jistý návod a doporučení, jak k těmto analýzám přistupovat, jaký by měl být postup a na co si dávat pozor. Domnívám se, že takový text v současnosti chybí. Na druhou stranu omezený rozsah příspěvku ve sborníku nedává dostatečný prostor na podrobnější popis a zejména dostatek ukázek.

Některé postupy jsou demonstrovány na 2 případových studiích: míra nezaměstnanosti v ČR a distribuce tweetů k dopravě metrem v metropolitní oblasti Madridu.

2. STRUČNÝ PŘEHLED PROSTOROVÝCH REGRESNÍCH MODELŮ

Pro teoretické základy lze odkázat na Anselin (1988, 2002), LeSage (1998), Haining (2003), Elhorst (2010), Smith et al. (2018), Ivan (2014), Horák (2019).

Základní otázky jsou, jak vyjádřit prostorovou závislost a jak ji efektivně integrovat do regresních rovnic.
Lze rozlišit dva rozdílné přístupy:

- prostorová závislost je řešena na **globální úrovni** s využitím explicitní podoby prostorové autokorelace (modely autoregresního typu jako je SAR či MRSA)
- prostorová závislost je řešena na **lokální úrovni** prostřednictvím lokálně proměnných parametrů modelu (např. GWR).

SAR (spatial autoregressive model) představuje v čisté podobě model, který používá k výpočtu cílové proměnné y pouze její hodnoty v sousedství.

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \boldsymbol{\varepsilon}$$

Autoregresní parametr ρ vyjadřuje prostorovou závislost a odhaduje se z dat s využitím zpravidla řádkové standardizace matice prostorových vah W (Smith et al, 2018); **ε** je chybový vektor.
Model v čisté podobě nevyužívá žádné nezávisle proměnné, proto se často o ně rozšiřuje a dostáváme smíšený regresně-prostorový autoregresní model (*mixed regressive spatial autoregressive* model *mrsa).* Patří k nim i jednoduché modely implementované v programu GeoDa, tj. spatial lag model SLM a spatial error model SEM.
Model prostorového kroku SLM specifikuje prostorovou závislost pouze pro cílovou proměnnou a ne pro nezávisle proměnné:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\rho W y} + \boldsymbol{\varepsilon}$$

s vektorem nezávisle proměnných X a jejich regresními koeficienty β, prostorovým autoregresním parametrem **ρ**, maticí vah **W** a chybovým vektorem **ε.**
Autokorelační chybový model SEM naopak vyjadřuje prostorovou závislost přes autokorelaci reziduí („chyb“) v modelu (Anselin, 2002):

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \quad kde\ \boldsymbol{\varepsilon} = \lambda\boldsymbol{W\varepsilon} + \boldsymbol{u}$$

kde λ je koeficient prostorové korelace chyb modelu, ε je vektor prostorově autokorelovaných chyb a u je vektor chyb.
Při optimalizaci se pak hledá, který z modelů SLM a SEM je v dané situaci kvalitnější. Uvádí se, že SLM se prosazuje tam, kde je menší vliv neznámých proměnných, zatímco pokud poskytuje lepší výsledky SEM, ukazuje to na významný vliv neznámých proměnných.
Podstatné je, že u **autoregresních modelů se získává 1 typ vztahu (rovnice) platná pro celé území**, tedy prostorové koeficienty ρ či lambda vyjadřují jistou průměrnou prostorovou závislost v celém území.
Odlišný přístup je použit u geograficky vážené regrese (GWR), kde se vztahy mezi proměnnými (rovnice) v území mění (Brunsdon et al., 1996). Metoda GWR tedy předpokládá **možnost existence prostorových odlišností ve vztazích** dvou a více proměnných a poskytuje způsob, jak tyto odchylky měřit.
Základem je vztah (Smith et al., 2018):

$$\mathbf{y} = \mathbf{X\beta(t)} + \boldsymbol{\varepsilon}$$

Regresní koeficienty β(t) jsou určovány ze sady bodů v definovaném okolí každého měřeného místa. Okolí je většinou kruhové s poloměrem r (anizotropní modely nejsou v současnosti podporovány). Namísto konstantní hodnoty r se používá vzdálenostní funkce vedoucí ke kernelovým odhadům
Dosah okolí se optimalizuje pomocí Akaike informačního kritéria (AIC) nebo křížovou validací. To však někdy selhává - u malého rozsahu to může skončit zahrnutím celého území nebo optimalizace v programovém prostředí (např. v ArcGIS) vůbec nelze řešit.
Výsledné odhady lokálních regresních parametrů se vizualizují a hodnotí z hlediska prostorové změny vztahu.
Při použití GWR tak můžeme získat jiný funkční vztah mezi dvěma proměnnými v závislosti na charakteru určitého regionu atd. Například vztah mezi cenou bytu a hustotou zalidnění v jeho okolí může být prostorově velmi odlišný, nebot' na cenu bytu má vliv množství dalších charakteristik, které uvedený vztah modifikují (Spurná 2008).
Doporučuje se provádět současně modelování oběma způsoby (autoregresní i lokálně regresní modely), protože každý z nich sleduje jiné vlivy a má jinou interpretaci (Horák, Orlíková 2019).

3. POSTUP PŘÍPRAVY MODELU

Na začátku přípravy každého regresního modelu je potřebné si dobře rozmyslet, co je závisle proměnná Y, jaké jsou nezávisle proměnné X_1 až X_n, resp. vysvětlující proměnné.

To nemusí být tak jednoduché, jak vypadá, viz další text k provedení EDA, kdy se zvažuje, které proměnné a v jaké podobě se použijí v regresním vztahu.

Klíčové je také rozhodnutí, jaký typ modelu připravujeme. Zjednodušeně lze říci, že rozlišujeme:

- Exploratorní (vysvětlující) s pevně danými proměnnými, u nichž nás zajímají jejich vztahy a vliv na Y
- Prediktivní (předpovědní) kde je cílem je co nejlépe vypočítat (odhadnout) Y ze sady proměnných.

Na začátku každého modelování se používá průzkumová analýza dat (Exploratory Data Analysis EDA), jejímž cílem je poznat vlastnosti datových sad. U prostorových dat se pak více mluví o ESDA (Exploratory Spatial Data Analysis), která zkoumá i prostorové vlastnosti dat.

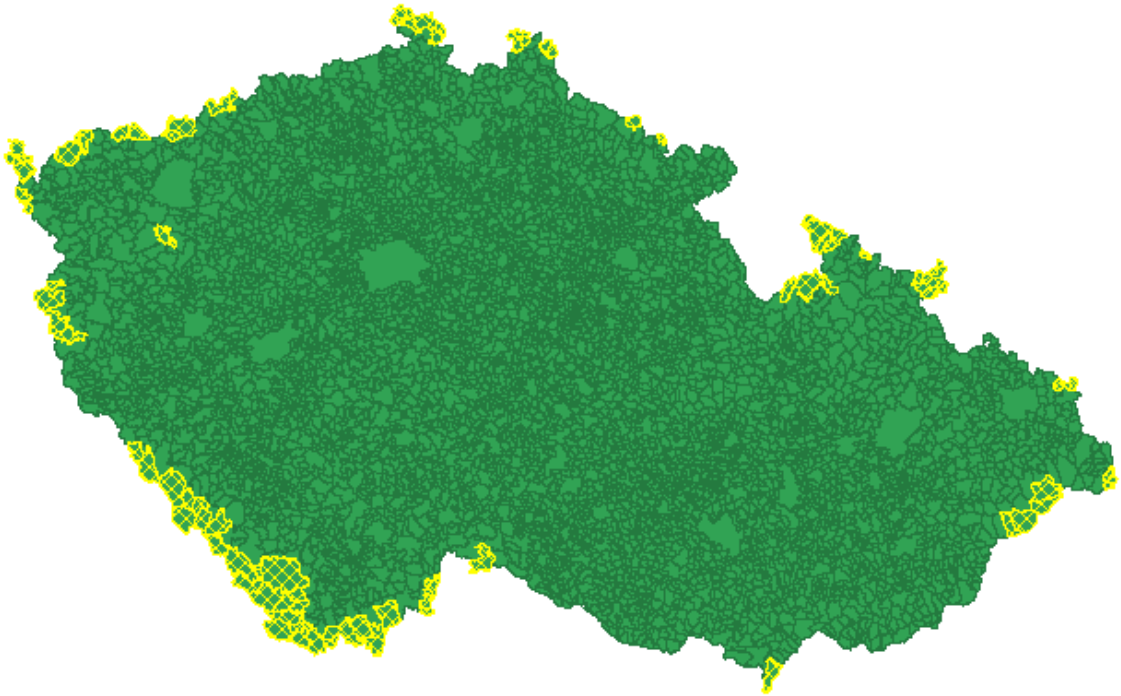
EDA zahrnuje analýzu distribuce každé proměnné, zejména ocenění její asymetrie a provedení vzájemné korelační a regresní analýzy všech proměnných. Cílem EDA je odhalit problémy proměnných s heteroskedascitou a odstranit multikolinearitu vznikající díky úzkým korelacím nezávisle proměnných.

Je vhodné také počítat sekundární proměnné, jako jsou poměry (míry, kvocienty, indexy) a kvůli problému MAUP (Openshaw, 1984) také hustoty (přepočty na plochu).

Pro odstranění asymetrie je zejména u predikčních modelů potřebné provést vhodnou transformaci, která zlepší symetrii distribuce proměnné, jinak mohou být vychýlené výsledné odhady závisle proměnné. U vysvětlujících modelů se ale po transformaci můžeme potýkat s vhodnou interpretací, protože daleko snáze se interpretují původní veličiny (zejména s využitím jejich standardizovaných regresních koeficientů).

Pro predikční modely, zejména u proměnných s velmi odlišnými rozsahy hodnot, je zpravidla potřebné provést standardizaci proměnných, nejběžnější je Z-standardizace. Opět platí, že standardizace v případě vysvětlujících modelů, může komplikovat interpretaci vlivu jednotlivých nezávisle proměnných.

Z hlediska **ESDA** je třeba správně volit typ prostorového sousedství a parametry (typy sousedství viz např. Horák, 2019). Je nutné zvažovat, jaký princip sousedství je v dané prostorové situaci vhodný a jaké bude mít volba důsledky. Nelze spoléhat slepě na optimalizační postupy. Je potřebné kontrolovat počty sousedů a všímat si zvláštních situací (Praha, okrajová místa apod.). Jednoduché kontroly nabízí např. GeoDa (viz obr. 1, kdy po nastavení sousedství na Euklidovskou vzdálenost 15 km se zjišťuje, které obce budou ovlivněny malým počtem dat a kolik jich bude).



Obr. 1 Obce ČR s méně než 15 sousedy do vzdálenosti 15 km v programu GeoDa

4. OLS

Na začátku prostorového regresního modelování se provádí standardní mnohonásobné lineární regresní modelování optimalizované zpravidla metodou nejmenších čtverců, proto se používá zkratka OLS (Ordinary Least Squarre). OLS umožňuje získat představu o celkovém (globálním) chování proměnných v regresní rovnici. Představuje i určitý benchmark, protože s jeho statistickými výsledky (oceňujícími kvalitu modelu) se porovnávají výsledky prostorového regresního modelování a pochopitelně se předpokládá, že prostorový regresní model je musí výrazně zlepšit, jinak lze těžko prokázat přidanou hodnotu prostorového regresního modelu.

Předpoklady dat pro mnohonásobnou lineární regresní analýzu (upraveno podle Vaus, 2002, Rabušic et al., 2019):

1. Závisle proměnná Y musí být metrická proměnná (numerická v kontinuální škále, alespoň intervalová data), jinak se musí provést logistická regrese.
2. Nezávisle proměnné mají být rovněž stejného typu, ale přípustná jsou i dichotomická (binární) data. U ostatních proměnných používáme umělé (dummy) proměnné.
3. Nezávisle proměnné nemají být mezi sebou vysoce korelované, aby nebyla v modelu multikolinearita
4. V datech nemají být odlehlé hodnoty, protože na ty je regresní analýza citlivá
5. Proměnné musejí být v lineárním vztahu, protože na tom je mnohonásobná lineární regresní analýza založena.
6. Vztahy mezi proměnnými nevykazují heteroskedascitu, tj. rozptyl veličiny se nemění v závislosti na hodnotě.
7. Počet záznamů (počet případů) musí být dostatečně velký vzhledem k počtu nezávisle proměnných. Doporučuje se 20 či alespoň 15 na každou nezávisle proměnnou. Uplatnění pravidla však není slepé. Pokud máte sice malý počet záznamů, ale každý reprezentuje agregaci např. z mnoha set či tisíc individuální případů), je to vyhovující.

Vztahy zejména mezi nezávisle proměnnými se prověřují korelační a regresní analýzou, v jejímž rámci se samozřejmě používají i neparametrické korelační koeficienty (např. Spearman, Kendal Tau) pro zachycení nelineárních vazeb. Kvůli multikolinearitě se doporučuje elimininovat proměnné s korelačním koeficientem vůči jiné nezávisle proměnné vyšším než 0.8 (Smith et al., 2018, Rabušic et al., 2019).

Pro posouzení významu jednotlivých proměnných v regresní rovnici a eliminaci případné multikolinearity se zpravidla používá Variance Inflation Factor (VIF). VIF jedné proměnné by měl být obecně nižší než 5 (Akinwande et al. 2015, Rabušic et al., 2019), existují ale i jiné názory na tuto hranici (např. 8). Vedle VIF se používá i tolerance, která má být menší než 0.2 (Rabušic et al., 2019).

Pro podrobnější vysvětlení mnohonásobné lineární regrese a příklady doporučuji knihu Rabušic et al. (2019).

Citlivou otázkou je vypuštění odlehlých hodnot, tedy anomálií z modelu. Odlehlé hodnoty samozřejmě negativně ovlivňují nejen statistickou kvalitu modelu, ale mohou i pokřivit charakteristiku reálných vztahů popsaných v regresních rovnicích. Bohužel není jednoznačné doporučení, zda (a jaké) odlehlé hodnoty z modelu odstranit či ne. Obecně platí, že je to více vhodné u prediktivních modelů než u exploratorních. Pochopitelně u velkých objemů dat je vyloučení několika hodnot méně citlivé než u malého souboru dat.

4.1 KVALITA MODELU OLS

K hodnocení kvality celého modelu se používá několik charakteristik a testů.

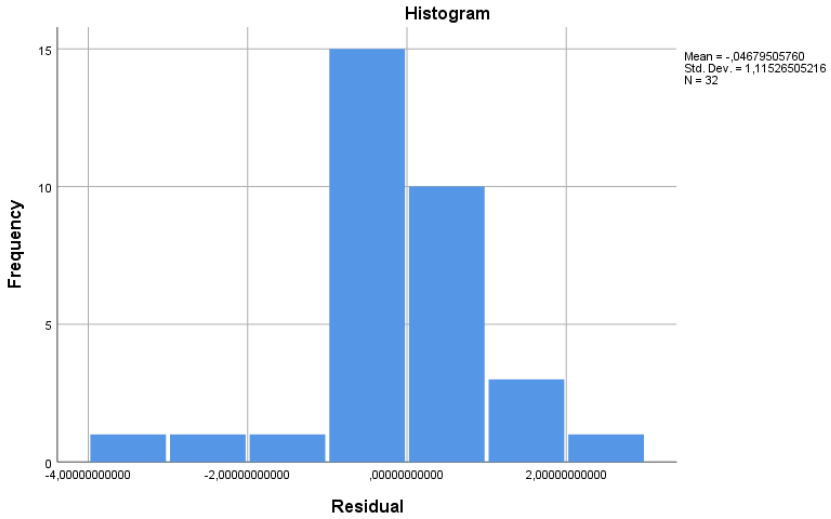
Na 1.místě se zpravidla uvádí koeficient determinace R^2. Je vhodné používat jeho adjustovanou variantu, zejména pro menší datové soubory (adjustovaný R^2 totiž přepočte R^2 na rozsah vzorku, což ukáže jeho reálnou sílu). Někdy tento přepočet může vést až k záporné hodnotě R^2, což samozřejmě neprospěje důvěře v kvalitu modelu. R^2 (adjustované) má být pochopitelně co největší, někteří autoři uvádějí, že alespoň 0.5 je potřebných pro kvalitní model. Přirozeně platí, že čím menší datový soubor a čím více nezávisle proměnných, tím snadněji získáte vysoké R^2. Proto zejména u malých datových souborů nespoléhejte pouze na R^2.

AIC (pozor na rozdíly v hodnotách AIC a AICc) by mělo být co nejmenší. CN (condition number) má být nízké, někteří autoři uvádějí do 30.

Testy heteroskedascity (Breusch-Pagan, Koenker-Bassett) mají potvrdit neexistenci heteroskedascity v datech.

Testy normality reziduí (Kolmogorov-Smirnov, Shapiro-Wilk, Jarque-Bera) mají potvrdit normalitu reziduí modelu. Jarque-Bera test se jeví jako jednodušší a snad i lépe splnitelný. Počítá se jednoduše z poměru mezi šikmostí a špičatostí distribuce (je využit např. v ArcGIS). Špatný výsledek jednoho z pokusných GWR

modelů pro madridská data je uveden na obr. 2 (histogram) a 3 (testy normality). Je třeba říci, že na rozdíl od dřívějších názorů se v současnosti normalita reziduí u velkých souborů nepovažuje za vážný problém, protože vzhledem k citlivosti testů stačí několik výchylek, aby test neprošel. V případě pochybností je vhodné prozkoumat distribuci reziduí a nespoléhat pouze na výsledek testu.



Obr. 2 Histogram vychýlených reziduí jednoho s pokusných GWR modelů v Madridu

Tests of Normality

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Residual	,237	32	,000	,831	32	,000

a. Lilliefors Significance Correction

Obr. 3 Nesplněné výsledky testů normality reziduí jednoho s pokusných GWR modelů v Madridu

Významnost modelu se posuzuje pomocí F-testu pro ANOVA.

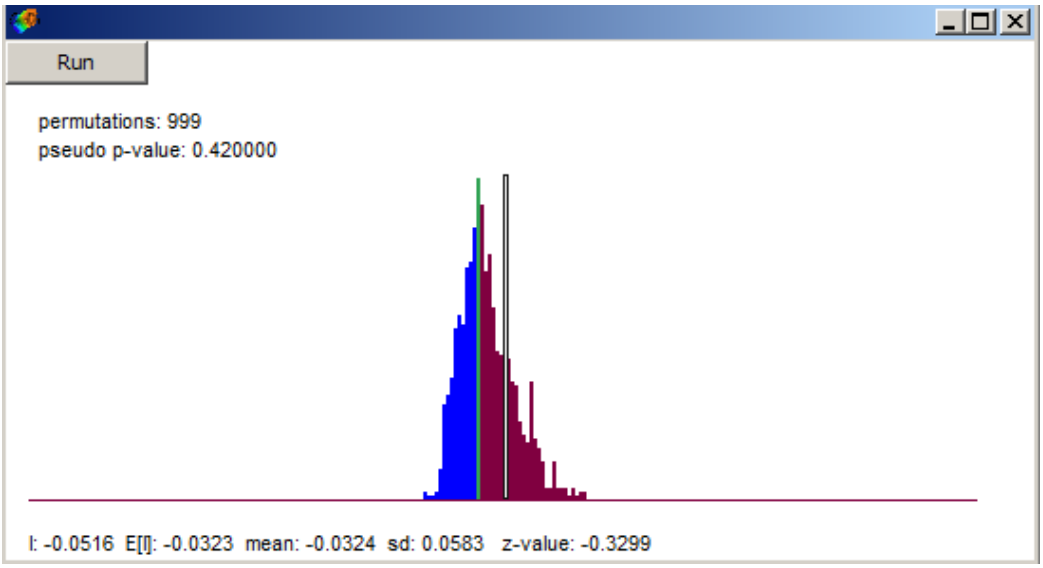
Testuje se rovněž prostorová autokorelace reziduí, k čemuž se často používá pouze jednoduchý výpočet Moranova I kritéria pro celou datovou sadu. Zatímco OLS model je v případě vysoké prostorové autokorelace reziduí špatný, je tato situace pro prostorové regresní modelování naopak výhodná, protože lze očekávat, že právě toto modelování problém odstraní a prostorovou autokorelaci naopak využije k podstatnému vylepšení modelu. Všimněme si, že třeba v ArcGIS průzkumném regresním modelování se jako výchozí hodnota pro výběr vhodného prostorového regresního modelu používá hodnota Moranova I > 0.5. Rovněž významnost prostorové autokorelace by měla být ověřena (např. permutacemi v prostředí GeoDa). Zatímco u OLS modelu často zjistíme významnou prostorovou autokorelaci reziduí, u výsledků prostorových modelů by měly být rezidua prostorově nekorelovaná (obr. 4).

Ve výsledném modelu se hodnotí také velikost absolutního členu v regresní rovnici (interecept) (konstanta v OLS či SAR). Pokud je příliš velký ve vztahu k hodnotám členů v rovnici, ukazuje to spíše na malou vysvětlující sílu modelu. Takový model pak může být jen matematickým řešením konkrétní úlohy, ale nepřispěje reálně k pochopení vazeb. Projevuje se to zejména u proměnných s asymetrickou distribucí a nelineárních vztahů.

ArcGIS nabízí exploratorní regresní modelování, které ze zadaného širokého seznamu nezávisle proměnných kombinatorickým způsobem hledá zajímavé regresní modely, které splňují nastavená kritéria kvality modelu. Příklad takového výstupu je zkráceně dokumentován v příloze 1.

Dobrý regresní model OLS je dokumentován v příloze 2 (výsledky z programu GeoDa) nebo v příloze 3 (výsledek jiného modelu v ArcGIS).

Pro interpretaci výsledků OLS modelu je vhodné také vypočítat standardizované beta koeficienty (viz https://www.statisticshowto.datasciencecentral.com/standardized-beta-coefficient/).



Obr. 4 Negativní prověření významnosti prostorové autokorelace reziduí modelu permutacemi v prostředí GeoDa (GWR model g2dens)

5. PŘÍPADOVÁ STUDIE MÍRY NEZAMĚSTNANOSTI V OBCÍCH ČESKA

Studie byla publikována v Horák, Orlíková (2019), zde jsou uvedeny jen vybrané části.

Cílem studie bylo posoudit významnost vlivu prostorového faktoru v regresním modelu nezaměstnanosti. Analýza byla provedena pro obce, protože se předpokládaly lepší prostorové vazby (vyšší prostorová autokorelace) i při vědomí komplikace ve značných rozdílech ve velikosti a tvaru obcí. Posuzovala se situace v březnu 2011 kvůli časovému souladu s daty SLDB 2011.

Cílovou proměnnou byla míra nezaměstnanosti, nezávisle proměnné se vybíraly z následujícího seznamu:

- Podíl populace ve věku 65+ (Age6500)
- podíl obyvatel se základním vzděláním a nižším (EduLow)
- podíl obyvatel s vysokoškolským vzděláním (EduUniv)
- Podíl věřících (Relig)
- Podíl cizinců (FOREIGN)
- Podilů rodáků (osoby narozené ve stejné obci) (NATIVE)
- Podíl dvou a více hospodařících domácností v 1 bytové domácnosti (TooHousH),
- Podíl obydlených bytů v bytových domech se sníženou kvalitou (PoorFlat),
- podíl hospodařících domácností v bytech s 5 a více členy domácnost (LargeFam),
- Podíl osob vyjíždějících do práce (Commute),
- Podíl osob denně vyjíždějících do práce (commuteD),
- Podíl zaměstnaných osob z 1000 obyvatel (EMPLO)
- Podíl volných míst na 1000 obyvatel (VACANT)

Po provedení EDA a ESDA se připravoval OLS model. Během vývoje byly vypuštěny proměnné TOOHOUSH (p=0.167) a COMMUTEND (p=0.143). Výsledné R^2 dosáhlo pouze 0.22. Vliv nezávisle proměnných odpovídal očekávání (obr. 5). Model však nebyl uspokojivý – hlavní problémem byla globální multikolinearita (CN=35), distribuce reziduí nebyla normální a existovala vysoká prostorová autokorelace reziduí.

Variables	Coefficient	StdError	t-stat	p	Robust_t	Robust_P	VIF
Intercept	13.058520	0.669	19.53	0.000*	15.1	0.000*	----
AGE6500	-0.215940	0.016	-13.32	0.000*	-9.9	0.000*	1.27
EDULOW	0.328725	0.016	20.20	0.000*	13.6	0.000*	1.69
EDUUNIV	-0.152825	0.025	-6.18	0.000*	-5.5	0.000*	1.69
RELIG	0.057473	0.007	7.89	0.000*	6.9	0.000*	2.18
FOREIGN	-0.107717	0.054	-1.98	0.048*	-2.7	0.006*	1.04
NATIVE	-0.099488	0.009	-10.68	0.000*	-8.6	0.000*	2.11
POORFLAT	0.234455	0.019	12.04	0.000*	8.7	0.000*	1.10
LARGEFAM	0.106314	0.019	5.58	0.000*	4.2	0.000*	1.42
COMMUTED	-0.042991	0.005	-8.43	0.000*	-7.1	0.000*	1.39
EMPLO	-0.001496	0.000	-5.87	0.000*	-3.5	0.000*	1.21
VACANT	-0.019269	0.008	-2.35	0.019*	-2.3	0.018*	1.13

Obr. 5 Seznam členů mnohonásobného lineárního regresního modelu vč. nestandardizovaných koeficientů (Horák, Orlíková, 2019)

Následně byl vyvíjen prostorový autoregresní model. Byly porovnány modely SLM a SEM a vybrán SEM podle vyšších hodnot Lagrangeova multiplikátoru a Robust LM.

SEM model má R^2=0.40 a rovněž AIC je lepší než u OLS modelu. Výsledky modelu jsou uvedeny na obr. 6. Je zřejmá významná role prostorového autokorelačního faktoru λ.

Variables	Coefficient	Std.Error	z-value	Probability
CONSTANT	13.964	0.7925	17.61931	0.00000
AGE6500	-0.128	0.0157	-8.178051	0.00000
EDULOW	0.207	0.0154	13.4538	0.00000
EDUUNIV	-0.120	0.0238	-5.051099	0.00000
RELIG	-0.030	0.0088	-3.42988	0.00060
NATIVE	-0.097	0.0087	-11.14488	0.00000
POORFLAT	0.229	0.0176	12.99703	0.00000
LARGEFAM	0.088	0.0171	5.155249	0.00000
EMPLO	-0.0014	0.0002	-6.751055	0.00000
LAMBDA	0.8969	0.0156	57.60186	0.00000

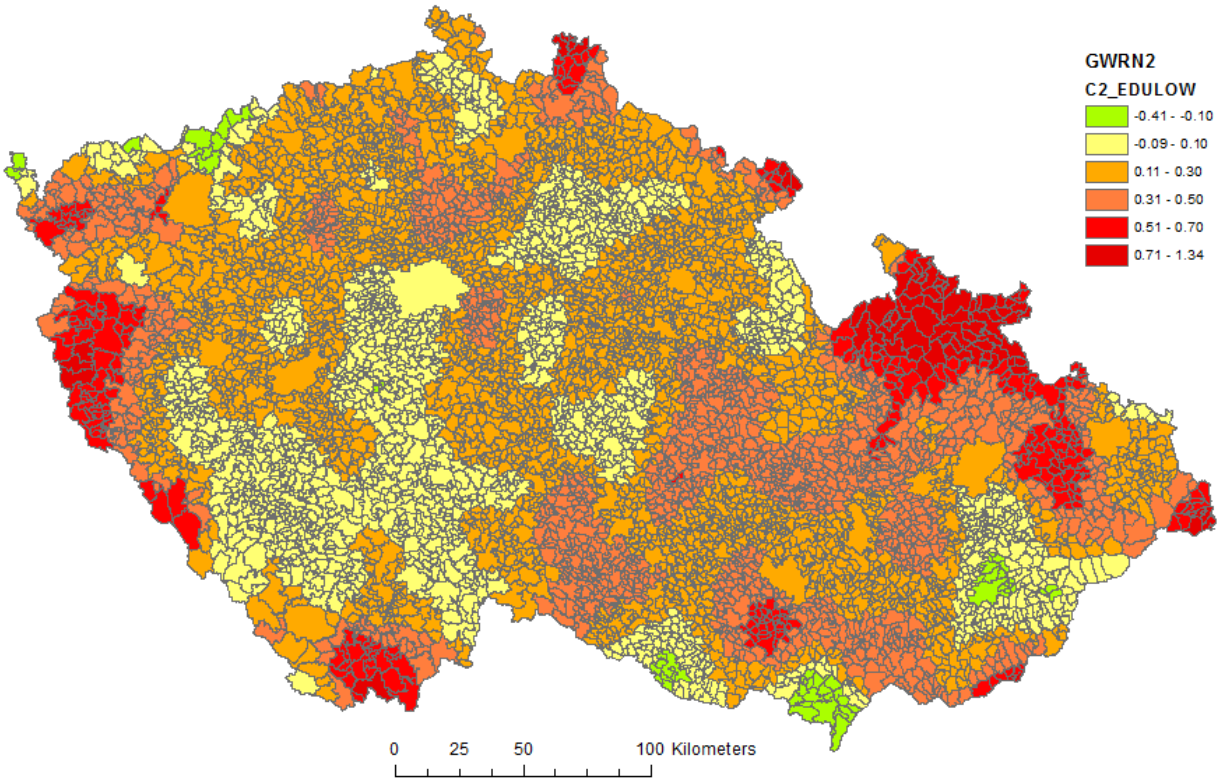
Obr. 6 Výsledky Spatial Error Modelu pro míru nezaměstnanosti v obcích ČR 3/2011 (Horák, Orlíková, 2019)

Nakonec byl vytvořen rovněž **geografický vážený model GWR**. Šířka pásma byla optimalizována podle AIC a adjusted R^2 na 15 km. Výsledný model je sice uspokojivý, přesto má nadále problémy s vysokou lokální multikolinearitou (vysoké CN, i když ESDA neukázala žádný problém) a podle očekávání distribuce rezudií není N (což je ale vzhledem k velikosti vzorku pochopitelné). Výsledná tabulka na obr. 7 ukazuje rozsah hodnot jednotlivých proměnných regresních koeficientů a podíly zastoupení kladných a záporných hodnot.

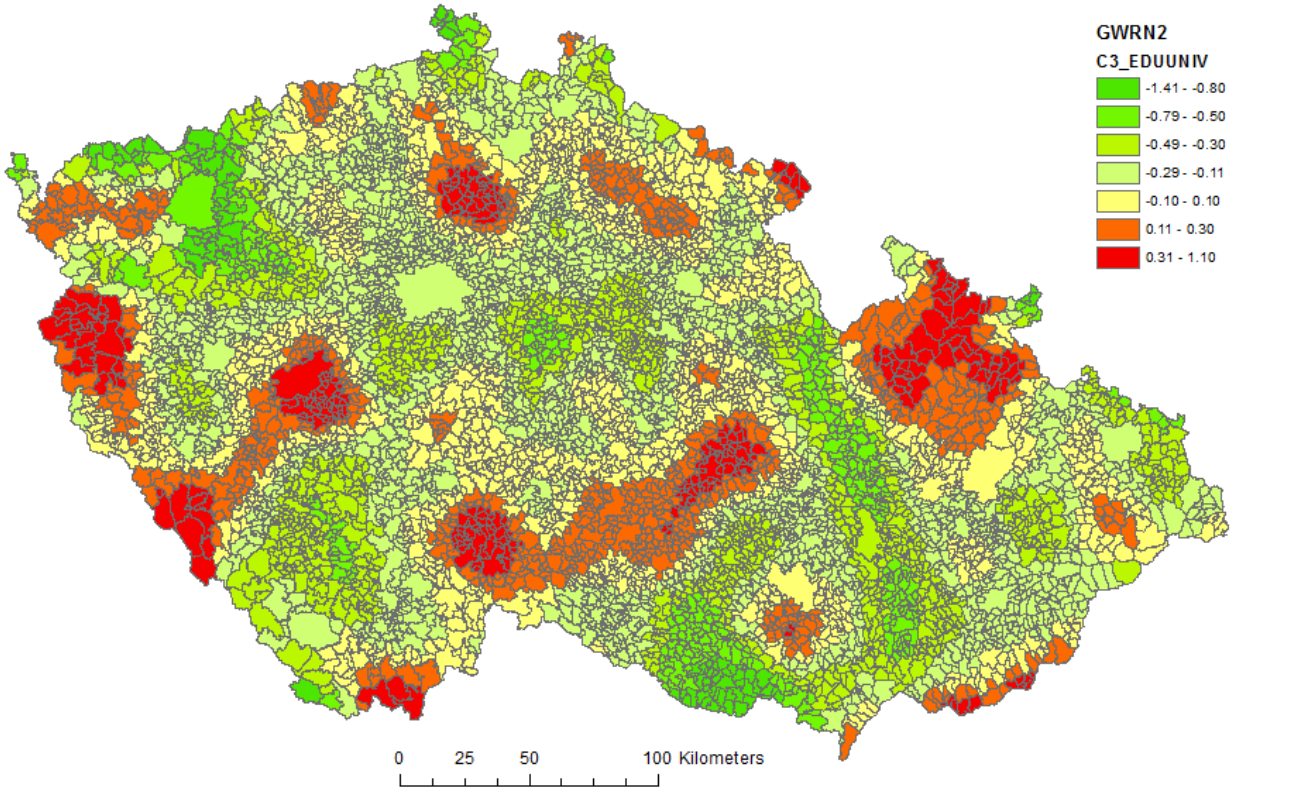
Variables	min	median	max	% of negative	% of positive
Intercept	-16.32	14.16	44.89	2.38	97.62
AGE6500	-0.84	-0.10	0.83	74.07	25.93
EDULOW	-0.43	0.2	1.34	9.16	90.84
EDUUNIV	-1.41	-0.16	1.10	74.00	26.00
RELIG	-0.91	-0.02	0.93	55.44	44.56
FOREIGN	-3.72	0.00	6.52	49.88	50.12
NATIVE	-0.44	-0.11	0.19	89.23	10.77
POORFLAT	-0.46	0.19	1.15	8.49	91.51
LARGEFAM	-0.58	0.06	1.13	33.92	66.08
COMMUTED	-0.44	-0.01	0.21	60.20	39.80
EMPLO	-0.04	-0.002	0.02	89.34	10.66
VACANT	-0.67	-0.002	1.34	50.92	49.08

Obr. 7 Výsledky GWR pro míru nezaměstnanosti v obcích ČR 3/2011 (Horák, Orlíková, 2019)

Pro ilustraci výsledků jsou uvedeny ještě 2 obrázky. První z nich (obr. 8) ukazuje hodnoty regresního koeficientu pro nezávisle proměnnou „podíl osob s nízkým vzděláním“, který ukazuje prakticky jednoznačně pozitivní vliv na míru nezaměstnanosti, avšak s rozdílnou silou tohoto vztahu. Vedle toho podíl osob s vysokoškolským vzděláním (obr. 9) vykazuje bipolární vliv, tj. jak očekávaný převládající negativní vliv na MN, tak i místa s pozitivním vlivem na MN (resp. kde převládají jiné faktory).



Obr. 8 Lokální regresní koeficienty vliv podílu osob s nízkým vzděláním na míru nezaměstnanosti v obcích ČR, 3/2011 (Horák, Orlíková, 2019)



Obr. 9 Lokální regresní koeficienty vlivu podílu osob s vysokoškolským vzděláním na míru nezaměstnanosti v obcích ČR, 3/2011

6. PŘÍPADOVÁ STUDIE DISTRIBUCE PŘÍSPĚVKŮ V SÍTI TWITTER K DOPRAVĚ METREM V MADRIDU

Vzhledem k tomu, že publikace hlavních výsledků regresního modelování teprve probíhá, omezíme se zde jen na základní informace a dokumentaci problémů, se kterými je možné se setkat při vývoji vhodného modelu.

Cílem studie bylo posoudit distribuci názorů lidí na veřejnou dopravu v Madridu a posoudit jejich příčiny. Komplikací při prostorovém hodnocení je zejména fakt, že tweety obsahují souřadnice pouze asi u 1% případů.

Pro potřeby analýzy byl použit dvouměsíční vzorek tweetů na oficiálním účtu metra. Ke stahování bylo využito Twitter API s vyloučením retweetů. Následovalo zpracování textu tweetů, jejich čištění (např. převod na malá písmena, transformace španělských znaků s diakritikou, odstranění speciálních znaků), geokódování a nakonec sémantická analýza a analýza sentimentu.

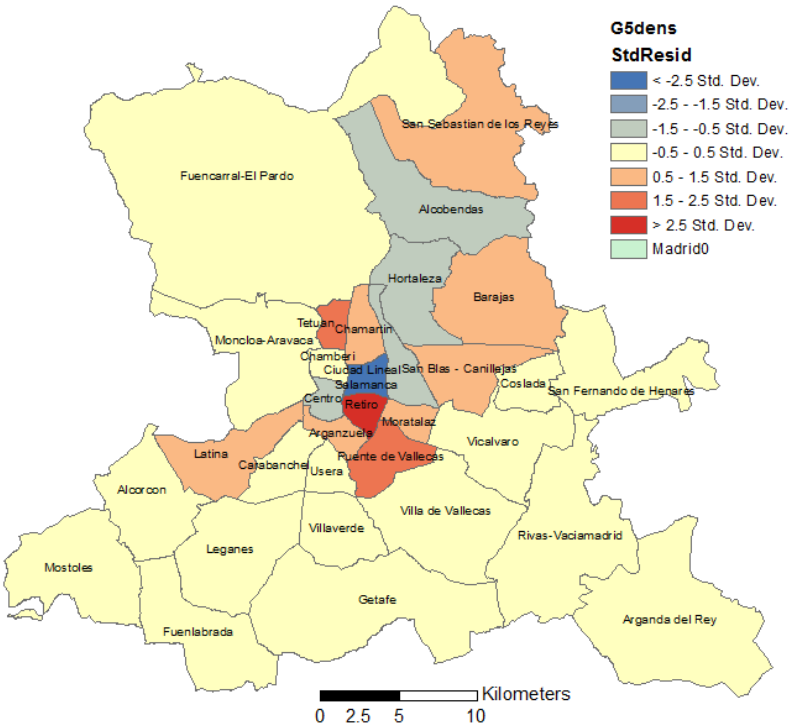
Při sémantické analýze se nejdříve zkoušela práce se slovníky a kombinacemi slov, která však dosáhla pouze cca 50% přesnosti. Následně byl použit jiný přístup, využívající shlukování textu pomocí Latent Dirichlet Allocation. Tato metoda dokázala odlišit 5 shluků, z nichž 4 poskytly vhodnou interpretovatelnost. Jednotlivé shluky tak reprezentují stížnosti osob na přesnost, komfort, poruchy a přeplnění dopravy. Celkově se podařilo dosáhnout cca 70% přesnosti.

Při analýze sentimentu bylo přiděleno pozitivní nebo negativní skóre každému tweetu. Byl využit BERT model hlubokého učení (Devlin, Chang, Lee, & Toutanova, 2018), který na rozsáhlé datové bázi dosáhl přesnosti cca 90%. Následně byla provedena klasifikace tweetů z dvouměsíční datové báze.

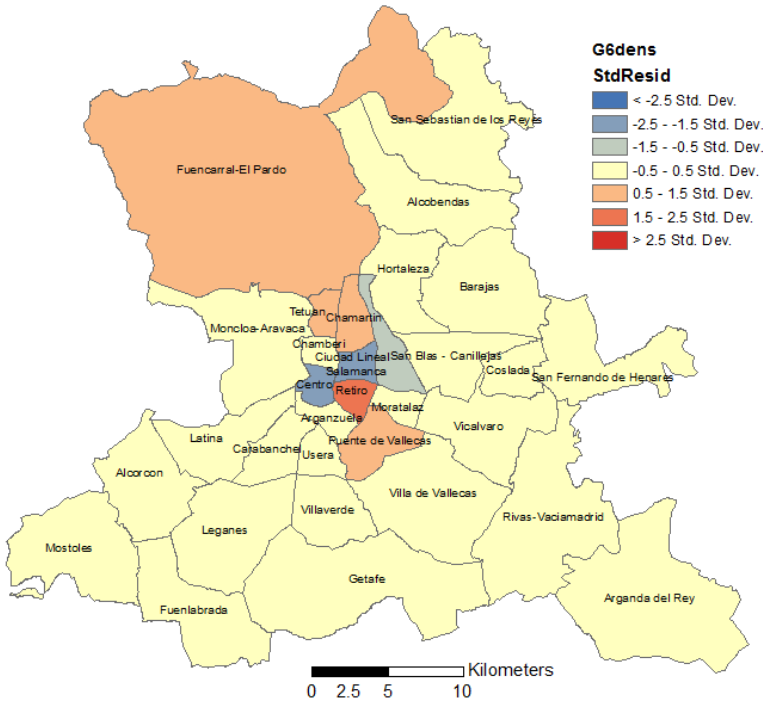
Z počtu tweetů byly odvozeny vhodné sekundární proměnné (hustoty, podíly, transformované a standardizované hodnoty) a podobně byly variantně zpracovány i nezávisle proměnné (obyvatelstvo, příjem, časový interval mezi spoji, POI, stanice, možnosti přestupů). Výsledky byly analyzovány pomocí EDA, ESDA, modelů OLS a GWR.

Při volbě sousedství se ukázalo, že nelze použít implicitní optimalizace prostorových vah, která vybírala všech 32 jednotek, a na základě logické úvahy byl vybrán model k-sousedů, po krokové optimalizaci bylo dosaženo k=12.

Při vývoji modelů se prokázalo jako užitečné sledovat standardizovaná rezidua modelu. Jak již bylo uvedeno, měla by být v absolutní hodnotě menší než 2,5. Na obr. 10 jsou vykreslena rezidua modelu G5dens, kde část Retiro přesahuje hodnotu 3 a Salamanca 2.8, což není akceptovatelné. Upravený model G6dens již má chyby akceptovatelně vysoké (obr. 11).



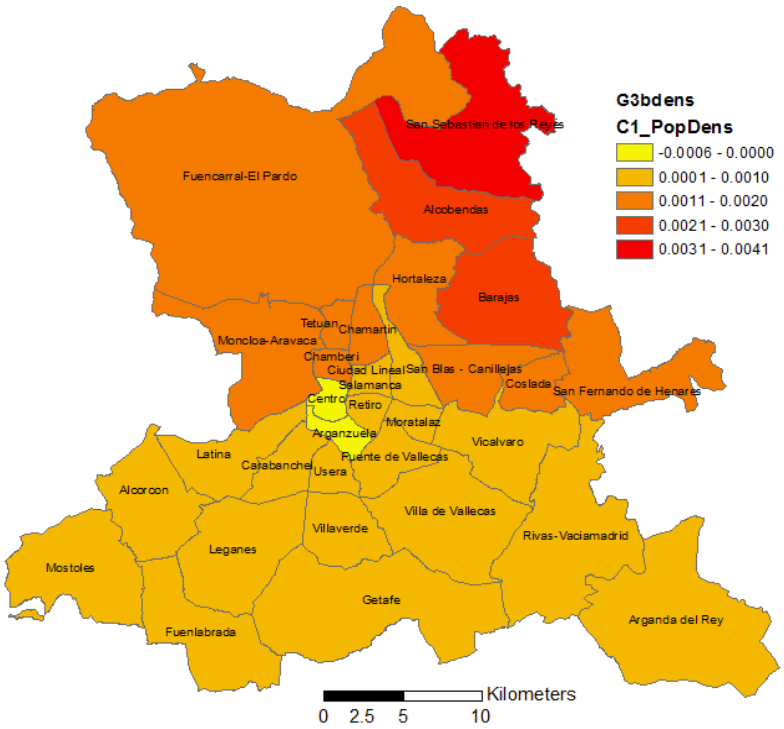
Obr. 10 Standardizovaná rezidua modelu G5dens



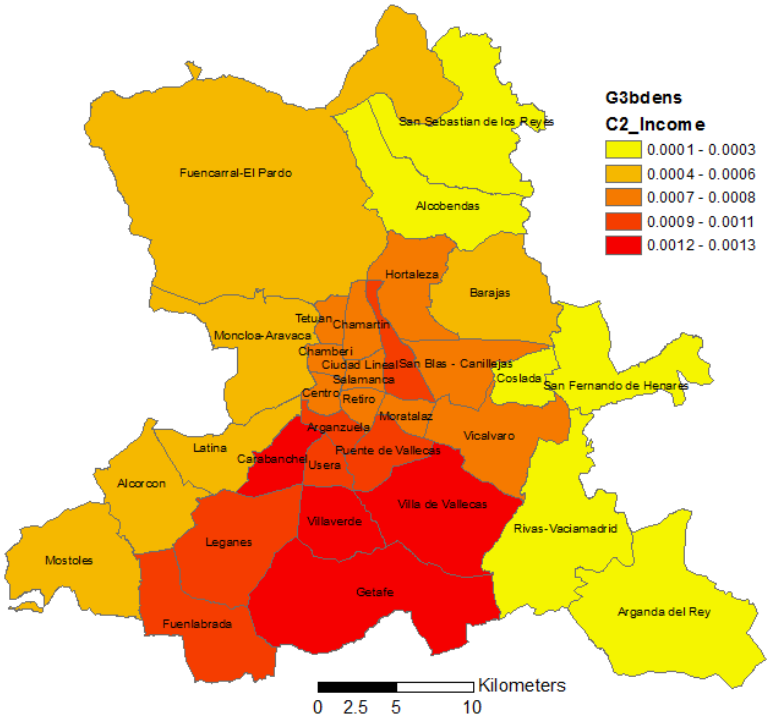
Obr. 11 Standardizovaná rezidua modelu G6dens

Jeden z výsledných modelů GWR sledoval jako závisle proměnnou hustotu stěžujících si uživatelů v závislosti na hustotě obyvatel, jejich příjmu a hustotě POI. Lokální regresní koeficienty pro hustotu obyvatel a pro příjem jsou uvedeny na následujících 2 obrázcích. Je zřejmé, že hustota obyvatel (obr. 12) zvyšuje hustotu stížností na sociální síti, tento vliv je však nejvýraznější v SV části Madridu, zatímco v centru je minimální až záporný. To je dáno skutečností, že v centru není hustota rezidenčních obyvatel významným

faktorem, ale prosazuje se mnohem více vliv hustoty POI, který odráží počty turistů, tvořících významnou část cestujících. Naproti tomu příjem obyvatel (obr. 13) zesiluje stížnosti osob na dopravu zejména v jižní části území.



Obr. 12 Lokální koeficienty pro hustotu obyvatel v modelu G3bdens



Obr. 13 Lokální koeficienty pro příjem obyvatel v modelu G3bdens

7. ZÁVĚR

Cílem příspěvku bylo zejména poukázat na doporučené postupy a některé problémy při prostorovém regresním modelování, s využitím dvou případových studií – míry nezaměstnanosti v Česku a distribuce tweetů k dopravě v Madridu.

Doporučený postup pro prostorové regresní modelování:

- Zvážit typ modelování (účel)
- Zvážit, jaké proměnné vybrat do regresního vztahu a v jaké formě (odlehlé hodnoty, transformace, standardizace, sekundární proměnné).
- Použít EDA a ESDA pro průzkum dat, kontrolu vztahů a studium odlehlých hodnot
- Vytvořit OLS model jako benchmark
- Zvážit typ sousedství a jeho nastavení (nepoužívat slepě optimalizace AIC)
- Pozor na velikost standardizovaných reziduí
- Vytvořit ideálně jak autoregresní tak lokální regresní model.
- Mapovat regresní koeficienty v GWR a provést jejich interpretaci

V každém případě je potřebné hodně přemýšlet o použitých datech, o výsledcích a kriticky zvažovat jejich interpretaci.

LITERATURA

1. Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. Open Journal of Statistics, 05(07), 754–767. https://doi.org/10.4236/ojs.2015.57075
2. Anselin L. (1988). *Spatial Econometrics: Methods and Models*. London: Kluwer. 284 s.
3. Anselin, L. (2002). Under the Hood: Issues in the Specification and Interpretation of Spatial Regression Models. *Agricultural Economics* 27(3). s. 247-267.
4. Brunsdon C, Fotheringham AS, Charlton M (1996) Geographically weighted regression: A method for exploring spatial nonstationarity. *Geogr Anal* 28(4). s. 281–298.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Dostupné z http://arxiv.org/abs/1810.04805
6. Elhorst, J.P. (2010). Spatial Panel Data Models. In Fischer, M.M. and Getis, A., Eds., *Handbook of Applied Spatial Analysis: Software Tools, Methods and Applications*, Springer Berlin Heidelberg, Berlin, 377-407. http://dx.doi.org/10.1007/978-3-642-03647-7_19
7. Haining, R. (2003). *Spatial Data Analysis Theory and Practice*. Cambridge University Press, Cambridge. 432 s.
8. Horák, J. (2019): Prostorové analýzy dat. Ostrava: VŠB-TU Ostrava. 181 s. ISBN 978-80-248-4368-1. Dostupné na http://homel.vsb.cz/~hor10/Vyuka/PAD/SkriptaPAD2019.pdf
9. Horák, J., Orlíková, L. (2019): Spatial Component in Regression Modelling of Unemployment in Czechia. In Proceedings of SMSIS (Strategic Management and its Support by Information Systems) 2019, Ostrava, 21-22.5.2019. 15 s.
10. Ivan (2014): *Kvantitativní metody v geografii*. VŠB-TU Ostrava, 2014. 110 s.
11. LeSage J. P. (1998). *Spatial Econometrics*. 273 s.
12. Openshaw, S. (1984). The modifiable areal unit problem. CATMOG (38). ISBN 0306-6142
13. Spurná P. (2006): Současné trendy v kvantitativní analýze geografických dat se zaměřením na využité metody geograficky vážené regrese DP Katedra sociální geografie a regionálního rozvoje PřF UK, Praha, 150 s.
14. Rabušic, L., Soukup, P., Mareš, P. (2019): Statistická analýza sociálněvědních dat (prostřednictvím SPSS). 2. přeprac.vyd. Brno: Masarykova univerzita. ISBN 978-80-210-9248-8.
15. Smith M.J., Goodchild M.F., Longley P.A.: Geospatial Analysis. http://www.spatialanalysisonline.com 2018.
16. Vaus, D. (2002): Analysing Social Science Data. London: Sage.

PŘÍLOHA 1 - ZKRÁCENÝ VÝPIS EXPLORATORNÍ REGRESNÍ MODELOVÁNÍ V ARCGIS PRO DATA TWITTER Z MADRIDU

```
Executing: ExploratoryRegression Madrid0 TOTNEGDENS
PopDens;Income;METRUSDENS;StopDens;POIdens;TimeInter;RStatTra;RStaTraNM # # # 8
1 0,5 0,05 7,5 0,1 0,1
Start Time: Wed Jan 15 12:19:19 2020
Running script ExploratoryRegression...
************************************************************************
Choose 1 of 8 Summary (výpis pro modely s 1 nezávisle proměnnou)
     Highest Adjusted R-Squared Results
AdjR2   AICc   JB K(BP)  VIF   SA   Model
 0,95 176,72 0,00  0,03 1,00 0,21  +METRUSDENS***
 0,92 194,18 0,08  0,00 1,00 0,67  +POIDENS***
 0,90 200,99 0,26  0,00 1,00 0,09  +STOPDENS***
       Passing Models
AdjR2 AICc JB K(BP) VIF SA   Model

(prázný výpis ukazuje, že žádný model v této kombinaci nesplnil nastavená
kritéria)

************************************************************************
Choose 2 of 8 Summary (výpis pro modely se 2 nezávisle proměnnými)
          Highest Adjusted R-Squared Results
AdjR2   AICc   JB K(BP)  VIF   SA   Model
 0,97 168,24 0,00  0,75 9,17 0,13  +METRUSDENS***  +POIDENS***
 0,95 177,46 0,00  0,12 1,00 0,27  +METRUSDENS***  +RSTATRANM
 0,95 179,03 0,00  0,06 1,63 0,25  +METRUSDENS***  +RSTATTRA
                                  Passing Models
AdjR2    AICc     JB    K(BP)   VIF    SA         Model
0,941    185,9    0,14  0,053   1,36   0,76  +POIDENS***  +RSTATTRA***
0,940    186,3    0,59  0,024   1,03   0,97  +INCOME***  +POIDENS***
0,935    189,1    0,47  0,108   2,49   0,95  +POPDENS***  +POIDENS***
0,905    201,1    0,27  0,001   1,01   0,11  +STOPDENS***  +RSTATRANM**
```

Atd.

Na konci je shrnutí modelování:

```
******** Exploratory Regression Global Summary (TOTNEGDENS) *************

              Percentage of Search Criteria Passed
                   Search Criterion Cutoff Trials # Passed % Passed
             Min Adjusted R-Squared > 0,50    255      244    95,69
            Max Coefficient p-value < 0,05    255       21     8,24
                      Max VIF Value < 7,50    255      119    46,67
            Min Jarque-Bera p-value > 0,10    255       53    20,78
Min Spatial Autocorrelation p-value > 0,10     33       32    96,97

------------------------------------------------------------------------------
             Summary of Variable Significance
Variable   % Significant % Negative % Positive
METRUSDENS        100,00       0,00     100,00
POIDENS           100,00       0,00     100,00
STOPDENS           37,50      32,81      67,19
RSTATTRA           25,00      25,00      75,00
POPDENS            22,66      33,59      66,41
INCOME             20,31      13,28      86,72
RSTATRANM          11,72      21,09      78,91
TIMEINTER           1,56      38,28      61,72
```

Summary of Multicollinearity

Variable	VIF	Violations	Covariates
POPDENS	7,44	0	--------
INCOME	2,25	0	--------
METRUSDENS	29,67	96	STOPDENS (98,46), POIDENS (98,46), RSTATTRA (24,62)
STOPDENS	33,50	94	METRUSDENS (98,46), POIDENS (90,77), RSTATTRA (24,62)
POIDENS	15,33	91	METRUSDENS (98,46), STOPDENS (90,77), RSTATTRA (24,62)
TIMEINTER	1,81	0	--------
RSTATTRA	10,08	30	STOPDENS (24,62), METRUSDENS (24,62), POIDENS (24,62)
RSTATRANM	5,18	0	--------

Summary of Residual Normality (JB)

JB	AdjR2	AICc	K(BP)	VIF	SA	Model
0,988859	0,952287	182,649849	0,147926	2,526571	0,751901	+POPDENS*** +INCOME*** +POIDENS*** +RSTATRANM
0,907390	0,956535	181,764998	0,340014	3,059726	0,933329	+POPDENS*** +INCOME*** +POIDENS*** +TIMEINTER* +RSTATRANM
0,873336	0,954101	181,409506	0,117773	2,771038	0,486932	+POPDENS** +INCOME** +POIDENS*** +RSTATTRA

Summary of Residual Spatial Autocorrelation (SA)

SA	AdjR2	AICc	JB	K(BP)	VIF	Model
0,990133	0,968799	167,168859	0,000000	0,185281	11,301509	+INCOME* +METRUSDENS*** +POIDENS***
0,971471	0,940156	186,307233	0,585753	0,023951	1,028238	+INCOME*** +POIDENS***
0,946378	0,934710	189,094161	0,470266	0,107793	2,486995	+POPDENS*** +POIDENS***

Table Abbreviations
AdjR2 Adjusted R-Squared
AICc Akaike's Information Criterion
JB Jarque-Bera p-value
K(BP) Koenker (BP) Statistic p-value
VIF Max Variance Inflation Factor
SA Global Moran's I p-value
Model Variable sign (+/-)
Model Variable significance (* = 0,10; ** = 0,05; *** = 0,01)

PŘÍLOHA 2 VÝSLEDEK RELATIVNĚ DOBRÉHO MODELU OLS V PROSTŘEDÍ GEODA

```
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  Madrid0
Dependent Variable  :  TOTNEGDENS  Number of Observations:   32
Mean dependent var  :     9.39425  Number of Variables   :    4
S.D. dependent var  :     16.2049  Degrees of Freedom    :   28

R-squared           :        0.956510  F-statistic           :     205.276
Adjusted R-squared  :    0.951850  Prob(F-statistic)     :3.67558e-019
Sum squared residual:     365.452  Log likelihood        :    -84.3724
Sigma-square        :     13.0519  Akaike info criterion :     176.745
S.E. of regression  :     3.61274  Schwarz criterion     :     182.608
Sigma-square ML     :     11.4204
S.E of regression ML:      3.3794

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT      -9.69656        2.30528       -4.20624     0.00024
           PopDens   0.000596339    0.000210264        2.83615     0.00839
            Income    0.00039049    0.000116043        3.36505     0.00223
           POIdens     0.0596027     0.00462762        12.8798     0.00000
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   8.977722
TEST ON NORMALITY OF ERRORS
TEST                   DF           VALUE              PROB
Jarque-Bera             2             0.6393          0.72640

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                   DF           VALUE              PROB
Breusch-Pagan test      3             4.8797          0.18082
Koenker-Bassett test    3             3.8588          0.27712
================== END OF REPORT ================================
```

PŘÍLOHA 3 VÝSLEDEK RELATIVNĚ DOBRÉHO MODELU OLS V PROSTŘEDÍ ARCGIS

Povšimněte si nevýznamnosti proměnných StopDens a TimeInter.

```
PopDens;Income;StopDens;POIdens;TimeInter
>>01/27/20 21:24:13
REGRESSION
----------
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            :  Madrid0
Dependent Variable  :  TOTNEGDENS  Number of Observations:   32
Mean dependent var  :     9.39425  Number of Variables   :    6
S.D. dependent var  :     16.2049  Degrees of Freedom    :   26

R-squared           :    0.960776  F-statistic           :     127.373
Adjusted R-squared  :    0.953233  Prob(F-statistic)     :1.99013e-017
Sum squared residual:     329.602  Log likelihood        :    -82.7204
Sigma-square        :      12.677  Akaike info criterion :     177.441
S.E. of regression  :     3.56048  Schwarz criterion     :     186.235
Sigma-square ML     :     10.3001
S.E of regression ML:     3.20937

------------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
------------------------------------------------------------------------------
          CONSTANT      -16.7216        6.04847        -2.7646     0.01034
           PopDens   0.000670545     0.00034947        1.91875     0.06606
           POIdens     0.0565493     0.00933369        6.05862     0.00000
            Income   0.000420885    0.000145265        2.89735     0.00754
          StopDens       1.06254        3.50805       0.302885     0.76439
         TimeInter       1.20074       0.797291        1.50602     0.14412
------------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   27.024993
TEST ON NORMALITY OF ERRORS
TEST                  DF           VALUE             PROB
Jarque-Bera            2             0.4312          0.80606

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF           VALUE             PROB
Breusch-Pagan test     5             8.8169          0.11659
Koenker-Bassett test   5             6.9205          0.22662
============================== END OF REPORT ================================
```