Reflection

LIANNE HANS
Supervisor: Dr. S. Koster
Faculty of Spatial Sciences, University of Groningen, The Netherlands.
E-mail: l.hans.1@student.rug.nl

Introduction

This document provides an additional reflection on the theory and methods used for the paper "The mediating role of settlement patterns on start-up activity in the urban-rural space". The first section reflects on the main focus of the paper and on the theories used. The second section reflects on the methodology and the last section discusses additional results not displayed in the paper.

Theoretical reflection

The aim of the paper is to explore the mediating effect of settlement patterns on the relationship between urbanization and start-up activity. Though the paper focuses on start-up activity, it fits in the wider debate concerning urban-rural relationships. The focus on start-up activity is motivated by the fact that start-ups are generally seen as an urban event: they are expected to be higher in more urbanized areas (see, for example, Van Oort & Atzema, 2004; Fritsch & Mueller, 2007; Bosma et al., 2008; Audretsch et al., 2015). Hence, theoretically, there is a distinctive urban-rural relationship for start-up rates. However, the empirical evidence on the role of urbanization in start-up patterns is ambiguous as some authors find that start-up rates are higher in more sparsely populated rural areas (Fritsch and Falck, 2007; Pettersson et al., 2010; Delfmann et al., 2014). In the paper, I argue that these ambiguous results may be explained by the fact that current studies have not explicitly taken into account the possible mediating effects of a country's level of urbanization and its settlement pattern on the relationship between urbanization and start-up activity. This is important to take into consideration, as for example a more rural area close to an urban center may have higher start-up rates than a more urbanized area located in a more peripheral region. While the paper focuses on start-up activity in studying the mediating effect of settlement patterns on the urban-rural relationship, I could also have focused on other factors with a distinctive urban-rural relationship, such as gross domestic product (GDP) per capita or employment. However, as start-ups are generally seen as generating economic growth (Audretsch & Keilbach, 2005; Bosma et al., 2011; Koster, 2011), it could be argued that patterns in GDP per capita and employment are to a large extent driven by start-up patterns.

The theoretical framework of the study is structured as follows: first it is explained why start-up rates are expected to be higher in urban areas; then I go on to discuss theories that might explain why this might not always be the case. Hence, the mediating factors in the urban-rural relationship. In the theoretical framework I combine theories from the entrepreneurship and innovation literature. The focus on theories from the innovation literature is motivated by the fact that new firm formation is generally seen as an innovative process (Kirchhoff et al., 2007; Baptista et al., 2008). Hence, there is a strong connection between innovation and start-up activity. However, I also discuss more general theories on urbanization economies and the benefits they provide for firms, as it is likely that regions

that provide these general benefits are also more conducive for start-up activity. Also, the second part also discusses theories from more general urban-rural studies, such as those from Partridge et al. (2007) on urban spillovers and Burger et al. (2015) on urban shadows and borrowed size. These theories provide more general arguments as to why urban-rural relationships may differ dependent on the settlement patterns of a country.

Alternatively to staying in the entrepreneurship literature, I could also have focused on more abstract theories, for example the rank-size rule and Zipf's law. Zipf's law states that the size of the n-ranked city is 1/n times the size of the largest city (Zipf, 1949). While this theory is useful to compare settlement patterns between countries it remains quite abstract. Moreover, Zipf's law focuses on cities and not so much on the urban-rural relationship. Hence, this theory is less useful in explaining differences in urban-rural relationships between countries. Hence, I would argue that the theories used in the paper are more insightful in explaining the main argument of the paper.

Methodological reflection

Cases and spatial unit of analysis

The empirical analysis aims at identifying the influence of the relative geography of a country, in terms of access to cities and the level of urbanization, on the spatial patterns of start-up intensity. For this purpose, the analyses focus on three European countries: The Netherlands, Belgium and Sweden. This is partly because of the good data availability in these countries and therefore it could be argued that the choice of countries is a bit arbitrary. However, these countries also provide especially interesting cases for the research, as the Netherlands and Belgium are among the most urbanized and densely populated countries of the OECD, while Sweden with its low population density and few large cities is a clear opposite. Hence, focusing on these three countries ensures that I have two clear opposite cases - Sweden and Belgium or Sweden and the Netherlands - and one similar case: Belgium and the Netherlands.

The spatial unit of analysis is the municipality. This low level of aggregation is needed as new firm formation is a local phenomenon (Sternberg, 2011; Audretsch et al., 2015). Indeed, most entrepreneurs start their business close to where they live (Figueiredo et al., 2002; Michelacci & Silva, 2007; Dahl & Sorenson, 2012), and a sizeable share of all start-ups operate from home (Mason et al., 2011). Hence, it is plausible to assume that entrepreneurs are mostly influenced by local conditions. However, the results indicate that the wider region also has an impact, implying that start-up rates are not only influenced by local conditions as is assumed in the literature (Sternberg, 2011; Audretsch et al., 2015). Yet, using a larger unit of analysis could have obscured the mediating effect of the level of urbanization of neighboring regions and distance to urban centers, as the effect diminishes with distance. Nevertheless, a problem with the unit of analysis is that the size of municipalities may differ between the three countries under consideration. However, partly this is controlled for by normalizing the variables; for example, I look at the number of new firm formations per 1000 people of working age as the dependent variable. Yet, with regard to the main explanatory variable, the level of urbanization, which is proxied by the population density, there remains a problem: for very large municipalities - mainly in the north of Sweden - the average population densities may be very low, even though most people may live concentrated in one place within the municipalities where population densities are higher. Therefore I excluded the north of Sweden as a

robustness check, the regression results of which are shown in appendix III of the paper; however, the results did not change significantly.

Using NUTS 3 as the spatial unit of analysis might have been more convenient as there is relatively good comparable data on this level available for most European countries so that the analyses could be done for the whole of Europe. Although one could argue that this would also reduce the problem of different definitions used as the data is retrieved from one platform, this platform - Eurostat - also receives its data from the national statistics offices of the different countries. Therefore, as most of the data for the analyses in the paper are from the national statistics offices, there should not be a large difference. Moreover, it could be argued that although the data for this study come from different sources, the availability of general guidelines provided by Eurostat contributes to an increasingly harmonized and synchronized data collection in the European Union (Audretsch et al., 2015); and this results in good comparability between the different national data sources. In addition, an important aim of the NUTS classification is to ensure that comparable regions appear at the same NUTS level. However, as population size has been defined in the Regulations as a key indicator for comparability, each level still contains regions that differ greatly in terms of area or economic weight (Eurostat, 2011). Thus, normalizing by population has a similar effect as using NUTS 3 as a level of analysis. Moreover, NUTS 3 level is too large for the analyses, due to the local nature of start-up activity (Sternberg, 2011; Audretsch et al., 2015). Another possibility was focusing on cities for the unit of analysis and comparing cities of similar size at different locations. However, urban spillovers to nearby rural areas cannot be considered in such an analysis. Though, one is able to study the mediating impact of the location of the city; hence, has Groningen a similar start-up level as Almere?

Main variables

The dependent variable is the rate of new firm formation. As the data do not allow distinguishing between genuinely new firm formations and new establishments of already existing firms, the research focuses at all new firm formations. It might have been better to consider solely genuinely new firm start-ups, as motivations and spatial patterns might be different for new subsidiaries (Van Oort & Stam, 2006; Koster, 2007). Moreover, the importance of urban proximity may also differ between the two types of start-ups (Koster, 2007). However, for both types of new firm formations it can be argued that urbanization and urban proximity is important as these areas provide the largest consumer markets (Stam, 2009; Bosma & Sternberg, 2014; Audretsch et al., 2015). The rate of new firm formation is calculated using the labor market approach. This approach uses the potential workforce in a region as the denominator for standardizing the number of new firm formations and is based on the assumption that each new firm is started by an individual person (Audretsch & Fritsch, 1994). The alternative, the ecological approach, uses the number of existing firms as the denominator, implying that new firms emerge from existing firms (Van Stel & Suddle, 2008). Using this measure can be misleading in areas with a small number of large firms (Garofoli, 1994).

The main explanatory variable is the level of urbanization. Population density is used as a proxy for urbanization, similar to earlier studies (see, for example, Verheul et al., 2002; Delfmann et al., 2014; Freire-Gibb & Nielsen, 2014; Audretsch et al., 2015). Alternatively, other measures could be used as a proxy for the level of urbanization, such as the number of firms per square kilometer or the labor market population per square kilometer. The first alternative however better fits the ecological approach, as it implicitly assumes that more existing firms result in more new firm formations. It can

be argued that the second alternatives better captures the possible supply of new entrepreneurs, as firms are generally started by economically active persons. However, using the total population per square kilometer ensures that both the demand and supply of new firm formations are included: although younger or older people might not be economically active, they still remain consumers.

Measuring the relative settlement patterns

In addition to the main explanatory variables, the analyses include variables that control for the mediating effect of settlement patterns. First, a spatially lagged variable is included, measuring the average level of urbanization - as measured by population density - in the surrounding municipalities. The spatial lag is calculated using a row-standardized spatial weights matrix based on inverse distances with a cut-off point of 50 kilometers. Hence, it is assumed that closer neighbors have a stronger influence on a municipality and that the impact of the level of urbanization of surrounding municipalities becomes zero after a distance of 50 kilometers. Alternatively, I could have used a matrices based on contiguity or a fixed distance band, where one imposes a "sphere of influence" onto the data: each feature is then analyzed within the context of those neighboring features within some specified critical distance and each neighboring feature has the same influence. However, using inverse distances is intuitively more appealing, as it is based on the first law of geography "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p. 236). The cut-off value is based on average and maximum commuting distances in the Netherlands, Belgium and Sweden, as most entrepreneurs start a new firm close to where they work or live (Figueiredo et al., 2002; Michelacci & Silva, 2007; Dahl & Sorenson, 2012). The value of 50 kilometers ensures that 90 per cent of all commuting travels are included (Sandow, 2008; Verhetsel et al., 2009; Statistics Netherlands, 2016). The other 10 per cent can be seen as extreme outliers. In addition, different cut-off points were used as a robustness check: 30 kilometers, 70 kilometers, and 100 kilometers. These results are shown in appendix I of the paper, but do not seem to be very sensitive to the cut-off value used.

I calculated the spatial lagged urbanization variable using Stata. Appendix I gives a copy of the Stata do-file I used to calculate the spatially lagged urbanization variable for the Netherlands. The process is similar for Belgium and Sweden. As a first step, I imported an ESRI shapefile including the x- and y-coordinates of the municipality centroids in Stata using the command "shp2dta" of Crow (2006). Subsequently, the spatial weights matrices are calculated using the command "spmat" command of Drukker et al. (2013). This command allows one to create, manage and store spatial-weighting matrices in Stata. Both contiguity and inverse-distance spatial-weighting matrices can be calculated using this command. For the contiguity spatial-weighting matrices, I had to manually adjust the spatial-weighting matrix for those municipalities without neighbors: i.e. islands. For this I exported the spatial-weighting matrix to a text-file using the "spmat export" command and adjusted the matrix so that the islands neighbor with the municipalities from which the boats to the island leave. After this I imported the matrices again in Stata. Using the command "spmat lag" I calculated the spatial lag of the urbanization variable. More precisely, "spmat lag" uses the spatial-weighting matrix calculated using "spmat" to compute the weighted averages of a variable, in this case the spatial lag of the urbanization (Drukker et al., 2013).

Initially, I also calculated a "relative geography variable" by subtracting the spatial lagged urbanization variable from the urbanization variable and dividing this value again by the spatially lagged urbanization variable. Hence, this "relative geography variable" would be the difference in

population density of a municipality with its neighboring municipalities relative to the average population density of these neighboring municipalities. Thus, this value indicated whether a municipality is relatively densely or relatively sparsely populated compared to its neighbors. However, in the end I chose to exclude this variable, due to two reasons. Firstly, the effect of this variable on start-up activity is not clear beforehand because it can take the same value for different situations. On the one hand, a high value may indicate that the municipality is relatively urbanized in an otherwise more rural region and this may positively influence start-up activity as there might not be very many alternative sources of agglomeration economies nearby. However, the effect on start-up activity might also be negative, as it indicates remoteness. On the other hand, the value may also be high for a highly urbanized municipality in an otherwise also urbanized region, and this may negatively impact start-up activity due to negative agglomeration economies. Also, if a municipality is relatively sparsely populated compared to its neighbors, this might positively influence start-up activity as this municipality may profit from the urbanization economies of its neighbors. These ambiguous effects are also reflected in the scatterplot in figure 1. More importantly however, the relative geography does not really add something after the spatial lag and the distance to urban centers variables - the distance variables will be discussed below - are included in the model. These variables already control for the mediating impact of the level of urbanization of the surrounding region and the location of a municipality relative to major settlements.

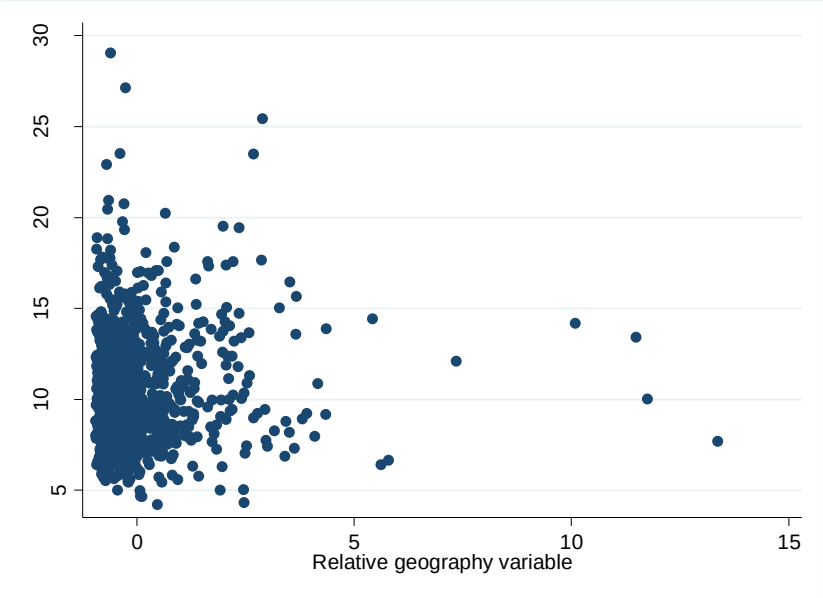

Figure 1. Scatterplot relating relative geography with start-up rates

To assess the impact of proximity and location in the urban system on the relationship between urbanization and start-up rates, the distance of a municipality to the nearest urban municipality, as well as to successively higher-tiered urban centers, is included in the analysis. First, I measured distance from the geographical centroid of the municipality to the centroid of the nearest urban municipality or the nearest urban municipality of a certain size. This is done in ArcGIS using the tool "near" from the proximity toolset. The distances are calculated using Euclidean distances. Although there may be measurement error bias when using straight-line distance rather than travel time, Apparicio et al. (2008) show that Cartesian distances (Euclidean and Manhattan distances) are strongly correlated with more accurate travel time distances. Moreover, Partridge et al. (2008) argue that such measurement error would bias the distance regression coefficient toward zero, suggesting

that the effect of distance would be stronger than reported. Also, with the relatively well-developed road systems in the countries under consideration, the measurement error is expected to be small.

Municipalities are defined as urban based on the "degree of urbanization" classification of Eurostat. Based on this classification, a municipality is seen as urban if fifty percent or more of the population lives in a high-density cluster, where high density clusters are defined as contiguous grid cells of one square kilometer with a density of at least 1 500 inhabitants per square kilometer and a minimum population of 50 000 (Dijkstra & Poelman, 2012). This definition is very useful, since it is based on grid cells which all have the same shape and surface thereby avoiding distortions caused by municipalities varying in size. The cut-off points for the different urban tiers are based on the OECD-EC definition (Dijkstra & Poelman, 2012), according to which small cities have between 50 000 and 100 000 inhabitants, medium cities have between 100 000 and 250 000 inhabitants and large cities have more than 250 000 inhabitants.

Alternatively, I could have looked at the centroids of the Urban Audit cities. However, the Urban Audit does not include all cities in a country. Moreover, the "degree of urbanization" classification for the municipalities is based on the same spatial unit as the rest of the analyses. Another option is to consider all cities in a country by regarding the x- and y-coordinates of the center of each city. For example, the distance to the nearest urban center for the municipality of Amsterdam would then be the distance from the centroid of the municipality to the center of the city of Amsterdam, which does not necessarily need to be zero. On the one hand this method corrects for the fact that a municipality can be classified by Eurostat as non-urban even if it does include a (small) city. This last is the case for Kiruna. On the other hand however, this method is based on a different spatial unit of analysis. Moreover, data for this method is not readily available and it is likely that mainly for the north of Sweden there is a problem that some municipalities are classified as non-urban while there is a major city in the municipality. However, as discussed before, the robustness check excluding the north of Sweden, reported in appendix III of the paper, does not seem to significantly alter the results.

Non-linear relationships

As it is argued in the introduction and theoretical framework that there might be some form of non-linearity in the relationship between urbanization and start-up activity, it might have been good to include a squared term of the urbanization variable in the analysis. Although I first did include a squared urbanization term, the results of this can be seen in table 1 below, I decided to not report these results in the final paper. The main reason for this was to not unnecessarily complicate the results as the inclusion of the squared term did not significantly change the other coefficients. Moreover, I used a logarithmic transformation for the urbanization variable; therefore, in a way, I already included some form of non-linearity. Also, the results seem a bit counter-intuitive: the effect of the squared urbanization term is significant and positive, whereas the effect on the original urbanization variable is negative. This implies that the effect of urbanization is negative, but that this negative effect is dampened after a certain point. This could off course be the case, in the sense that very sparsely populated areas experience a negative effect from urbanization because they lose for tourism valuable nature and open space, whereas after a certain point positive urbanization economies also become important. However, discussing this more in depth would deviate from the focus of the paper. To control for a possible "distance-protection" effect, the analyses do include the squared terms for the distance to urban centers variables. To reduce problems of multicollinearity, these variables are centered, i.e., the mean is subtracted from each value.

Maximum Likelihood Estimation

The main analysis performed is a linear regression. Hence, it assumed that there is a linear relationship between the dependent variable and the explanatory variables. As the scatterplot showed no clear linear relationship between the urbanization variable and start-up activity, the urbanization variable was log-transformed. Normally, linear regression models are estimated using Ordinary Least Squares (OLS). However, since OLS provides inconsistent standard errors for models including a spatially lagged explanatory variable (Gibbons & Overman, 2012), the linear model is estimated using Maximum Likelihood Estimation (MLE). A copy of the do-file used in Stata to estimate this model is displayed in appendix II.

An alternative to using a global model is doing a Geographically Weighted Regression (GWR). GWR allows relationships in a regression model to vary over space (Wheeler, 2014). In contrast to traditional linear models, GWR estimates regressions coefficients locally at spatially referenced data points. Hence, GWR is able to capture spatially varying relationships between covariates and an outcome variable (Wheeler, 2014). However, due to problems with multicollinearity, GWR is more appropriately seen as an exploratory tool and not as a formal model to infer parameter nonstationarity (Wheeler, 2014). Moreover, for my analysis GWR is less appropriate: GWR can indicate that the relationship differs between regions; however, it does not provide an explanation as to why this might be the case. Using a global model allows one to show the reader that the relationship becomes more similar for different countries after controlling for relative geography in terms of the level of urbanization of neighboring regions and the location relative to major urban areas.

Table 1. Regression results including the squared urbanization term: Maximum Likelihood Estimation for the Netherlands (NL), Belgium (BE) and Sweden (SE)

Dependent variable: start-up rate	NL		BE		SE	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Population density (ln)	-0.387***	-0.503***	-0.197***	-0.437***	0.0263	-0.556***
	(0.146)	(0.162)	(0.0655)	(0.104)	(0.0962)	(0.105)
Square of population density (ln)	0.0370	0.0834	0.206***	0.293***	0.0535**	0.0506**
	(0.0915)	(0.0880)	(0.0312)	(0.0363)	(0.0229)	(0.0219)
Spatial lag population density		-0.00228***		-0.000485***		0.000724
		(0.000459)		(0.000137)		(0.000470)
Dist to nearest UC		-0.0692***		-0.0134*		-0.00762***
		(0.0211)		(0.00746)		(0.00161)
Square of dist to nearest UC		0.00160*		-0.000261**		3.36e-05***
		(0.000938)		(0.000105)		(7.59e-06)
Inc Dist to UC >50 000		-0.178***		0.0177		0.283
		(0.0389)		(0.0591)		(0.260)
Square of Inc Dist to UC > 50 000		0.00194*		0.0147		-0.0243
		(0.00110)		(0.0124)		(0.0301)
Inc Dist to UC > 100 000		-0.0642***		-0.00950		0.0488
		(0.0167)		(0.0111)		(0.0673)
Square of Inc Dist to UC > 100 000		0.000832		-0.00102*		-0.000548
		(0.000586)		(0.000534)		(0.00334)
Inc Dist to UC > 250 000		-0.0381***		-0.0160***		-0.0112***
		(0.00534)		(0.00306)		(0.00130)
Square of Inc Dist to UC > 250 000		0.000122*		-0.000215***		1.53e-05***
		(6.45e-05)		(5.64e-05)		(3.51e-06)
Age < 15	0.344**	0.258**	-0.311***	-0.141	-0.00497	0.111
	(0.140)	(0.130)	(0.0916)	(0.0859)	(0.166)	(0.141)
Age 15-25	-0.0950	-0.0857	0.209***	0.339***	-0.0155	0.00935
	(0.114)	(0.105)	(0.0760)	(0.0746)	(0.141)	(0.122)

Age 25-35	-0.0539	-0.0328	0.178**	0.263***	-0.0937	0.135
	(0.139)	(0.133)	(0.0834)	(0.0782)	(0.117)	(0.103)
Age 35-50	*Reference*	*Reference*	*Reference*	*Reference*	*Reference*	*Reference*
Age 50-65	-0.205	-0.137	-0.0721	-0.0794	0.401***	0.419***
	(0.152)	(0.139)	(0.0878)	(0.0813)	(0.128)	(0.107)
Age 65+	0.150	0.266***	0.0679	0.195***	-0.00669	0.130
	(0.0918)	(0.0859)	(0.0565)	(0.0551)	(0.101)	(0.0872)
Higher educated	0.200***	0.150***	0.123***	0.128***	0.143***	0.160***
	(0.0219)	(0.0217)	(0.00998)	(0.0102)	(0.0267)	(0.0230)
Immigrant	0.0954***	0.0955***	0.000369	0.00103	-0.0108	-0.0379**
	(0.0221)	(0.0207)	(0.00372)	(0.00343)	(0.0164)	(0.0149)
Unemployment	0.100	0.132	-0.0188	0.00818	-0.0404	0.0119
	(0.126)	(0.116)	(0.0174)	(0.0179)	(0.0323)	(0.0278)
Service	0.0886***	0.0777***	0.00444	0.00842	0.180***	0.149***
	(0.0173)	(0.0159)	(0.00761)	(0.00716)	(0.0179)	(0.0155)
Public	0.0125	0.0114	-0.0481***	-0.0389***	-0.0625***	-0.0158
	(0.0154)	(0.0143)	(0.00663)	(0.00613)	(0.0227)	(0.0196)
Constant	-0.940	-0.145	7.561	-0.145	-1.723	-12.01
	(8.858)	(8.156)	(5.228)	(4.893)	(9.000)	(7.649)
Log Likelihood	-849.3274	-808.1563	-1001.109	-943.0634	-479.393	-417.5491
AIC	1726.655	1662.313	2030.219	1932.127	986.7859	881.0981
Wald chi2	390.65***	570.03***	448.08***	674.02***	596.17***	1067.53***
Observations	405	405	589	589	290	290

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Additional results

I already discussed why I did not include the squared urbanization variables or relative geography variables in the paper. The results for the analysis including the squared urbanization term are displayed in table 1. Both in this table and in tables in the paper the results are displayed separately for each country. Alternatively, I could have reported the results for the total sample, so for all countries together, as is displayed in table 2. Although this table gives some interesting results about the general relationship between start-up rates and the explanatory variables, the differences between the countries are obscured. Therefore, I chose to focus on the separate results for each country in the paper and compare them. Moreover, due to the different data sources used there are some small definition differences between the three countries which could cause distortions when including them all in one model.

Table 2. Regression results: Maximum Likelihood Estimation (MLE) for all countries

	All countries			
Dependent variable: start-up rate	Model 1	Model 2	Model 3	Model 4
Population density (ln)	-0.0470	-0.412***	0.00392	-0.344***
	(0.0547)	(0.0720)	(0.0566)	(0.0710)
Square of population density (ln)			0.0434***	0.134***
			(0.0131)	(0.0178)
Spatial lag population density		-0.000310***		-0.000728***
		(0.000101)		(0.000114)
Dist to nearest UC		-0.0159***		-0.0182***
		(0.00269)		(0.00264)
Square of dist to nearest UC		6.33e-05***		4.86e-05***
		(9.47e-06)		(9.47e-06)
Inc Dist to UC >50 000		0.0477**		0.0343
		(0.0243)		(0.0238)
Square of Inc Dist to UC > 50 000		-0.00248***		-0.00219***
		(0.000789)		(0.000773)
Inc Dist to UC > 100 000		-0.0233**		-0.0209**
		(0.0103)		(0.0101)
Square of Inc Dist to UC > 100 000		2.90e-05		-8.73e-05
		(0.000407)		(0.000398)
Inc Dist to UC > 250 000		-0.0172***		-0.0192***
		(0.00148)		(0.00147)
Square of Inc Dist to UC > 250 000		2.68e-05***		2.66e-05***
		(3.72e-06)		(3.64e-06)
Age < 15	-0.0315	0.0364	-0.0505	0.000695
	(0.0747)	(0.0708)	(0.0746)	(0.0695)
Age 15-25	-0.164***	-0.111*	-0.172***	-0.145***
	(0.0599)	(0.0567)	(0.0597)	(0.0557)
Age 25-35	0.0407	0.170***	-0.0170	0.0597
	(0.0663)	(0.0642)	(0.0683)	(0.0645)
Age 35-50	*Reference*	*Reference*	*Reference*	*Reference*
Age 50-65	-0.228***	-0.198***	-0.240***	-0.220***
	(0.0740)	(0.0697)	(0.0737)	(0.0682)
Age 65+	0.0760	0.195***	0.0370	0.120***

	(0.0468)	(0.0453)	(0.0481)	(0.0455)
Higher educated	0.147***	0.151***	0.146***	0.155***
	(0.00989)	(0.00949)	(0.00986)	(0.00930)
Immigrant	0.00720	0.00454	0.00738	0.00350
	(0.00452)	(0.00428)	(0.00451)	(0.00419)
Unemployment	-0.00907	0.0389**	-0.0155	0.0330**
	(0.0156)	(0.0154)	(0.0156)	(0.0151)
Service	0.0493***	0.0437***	0.0458***	0.0457***
	(0.00757)	(0.00733)	(0.00761)	(0.00718)
Public	-0.0198***	-0.0110*	-0.0212***	-0.0116*
	(0.00678)	(0.00642)	(0.00677)	(0.00628)
Sweden	-1.442***	-3.230***	-1.301***	-3.067***
	(0.417)	(0.423)	(0.417)	(0.415)
Netherlands	*Reference*	*Reference*	*Reference*	*Reference*
Belgium	-4.754***	-5.011***	-4.616***	-4.694***
	(0.186)	(0.178)	(0.190)	(0.179)
Constant	13.23***	7.032*	15.32***	10.78***
	(4.430)	(4.209)	(4.456)	(4.149)
Log Likelihood	-2568.185	-2481.923	-2562.747	-2454.175
AIC	5166.37	5011.847	5157.494	4958.351
Wald chi2	1916.92***	2377.25***	1944.15***	2538.96***
Observations	1 284	1 284	1 284	1 284

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The paper discusses two main models: one with no controls for settlement patterns and one that does control for relative geography by including the spatial lag and the distance to urban centers. However, I could also have displayed the results for the spatial lag and distance to urban centers separately. In that case I would have a model with no controls for geography, one which included the spatial lag, one which included the distance to urban centers and the final model that included both. However, I chose to not display the steps in between, as they did not provide much additional information for the point I wanted to make. The focus is on how controlling for geography impacts the relationship between urbanization and start-up activity and whether this is the same or different for each country. Hence, it is not particularly interesting whether this effect is mainly driven by the spatial lagged urbanization variable or the distance to the urban centers: both matter and therefore they should be considered in one model.

As a robustness check I also run the model for Belgium without Wallonia, and for Sweden without the north of Sweden. These results are displayed in the appendices of the paper. Alternatively, I could have included a dummy to control for the possible differing effects of Wallonia and North-Sweden. The Stata commands for this are included in appendix II. However, a dummy only adjusts the intercept, thereby implicitly assuming that the relationship between other explanatory variables and the dependent variable remain the same irrespective of whether the dummy variable takes a value of ‘1’ or ‘0’. Hence, for Belgium this would imply that the impact of urbanization on start-up activity is the same in Wallonia as it is in Flanders, although the general level of start-up activity may differ between the two regions. It is possible to create interaction variables of explanatory variables with the dummy variable to allow relationships to differ depending on the value of the dummy

variable; however, I think this would unnecessarily complicate the model, whereas excluding one of the regions illustrates the same point.

References

Apparicio, P., Abdelmajid, M., Riva, M. & Shearmur, R. (2008). Comparing alternative approaches to measuring the geographical accessibility of urban health services: distance types and aggregation-error issues. International Journal of Health Geographics, 7(1), 1-13.

Audretsch, D. B., Belitski, M., & Desai, S. (2015). Entrepreneurship and economic development in cities. The Annals of Regional Science, 55(1), 33-60.

Audretsch, D. B., & Fritsch, M. (1994). On the measurement of entry rates. Empirica, 21(1), 105-113.

Audretsch, D.B. & Keilbach, M. (2005). Entrepreneurship capital and regional growth. Annals of Regional Science, 39, 457-469

Baptista, R., Escária, V., & Madruga, P. (2008). Entrepreneurship, regional development and job creation: The case of Portugal. Small Business Economics, 30(1), 49–58.

Bosma, N., Stam, E. & Schutjens, V. (2011). Creative destruction and regional productivity growth: evidence from the Dutch manufacturing and services industries. Small Business Economics, 36, 401-418

Bosma, N., & Sternberg, R. (2014). Entrepreneurship as an urban event? Empirical evidence from European cities. Regional Studies, 48(6), 1016-1033.

Bosma, N., Van Stel, A. & Suddle, K. (2008). The geography of new firm formation: Evidence from independent start-ups and new subsidiaries in the Netherlands. International Entrepreneurship and Management Journal, 4, 129-146.

Burger, M.J., Meijers, E.J., Hoogerbrugge, M.M. & Masip Tresserra, J. (2015). Borrowed size, agglomeration shadows and cultural amenities in North-West Europe. European Planning Studies, 23(6), 1090-1109.

Crow, K. (2006). shp2dta: Stata module to convert shape boundary files to Stata datasets. Statistical Software Components S456718, Department of Economics, Boston College.

Dahl, M. S., & Sorenson, O. (2012). Home sweet home: Entrepreneurs' location choices and the performance of their ventures. Management Science, 58(6), 1059-1071.

Delfmann, H., Koster, S., McCann, P., & Van Dijk, J. (2014). Population change and new firm formation in urban and rural regions. Regional Studies, 48(6), 1034-1050.

Dijkstra, L. & Poelman, H. (2012). Cities in Europe: the new OECD-EC definition. Regional Focus 01/2012. Brussels: European Commission.

Drukker, D.M., Pehg, H., Prucha, I.R. & Raciborski, R. (2013). Creating and managing spatial-weighting matrices with the spmat command. The Stata Journal, 13(2), 242-286.

Eurostat (2011). Regions in the European Union. Nomenclature of territorial units for statistics NUTS 2010/EU-27. ISSN 1977-0375. Luxembourg: Publications Office of the European Union.

Figueiredo, O., Guimaraes, P., & Woodward, D. (2002). Home-field advantage: location decisions of Portuguese entrepreneurs. Journal of Urban Economics, 52(2), 341-361.

Freire-Gibb, L.C. & Nielsen, K. (2014). Entrepreneurship within urban and rural areas: creative people and social networks. Regional Studies, 48(1), 139-153.

Fritsch, M. & Falck, O. (2007). New business formation by industry over space and time: a multidimensional analysis. Regional Studies, 41(2), 157-172.

Fritsch, M. & Mueller, P. (2007). The persistence of regional new business formation-activity over time–assessing the potential of policy promotion programs. Journal of Evolutionary Economics, 17, 299-315.

Kirchhoff, B., Newbert, S., Hasan, I., & Armington, C. (2007). The influence of university R&D expenditures on new business formations and employment growth. Entrepreneurship: Theory & Practice, 31(4), 543–559.

Garofoli, G. (1994). New firm formation and regional development: the Italian case. Regional studies, 28(4), 381-393.

Gibbons, S., & Overman, H. G. (2012). Mostly pointless spatial econometrics? Journal of Regional Science, 52(2), 172-191.

Koster, S. (2007). The entrepreneurial and replication function of new firm formation. Tijdschrift voor Economische en Sociale Geografie, 98 (5), 667-674

Koster, S. (2011). Individual foundings and organizational foundings: their effect on employment growth in The Netherlands. Small Business Economics, 36, 485-501

Mason, C. M., Carter, S., & Tagg, S. (2011). Invisible businesses: the characteristics of home-based businesses in the United Kingdom. Regional Studies, 45(5), 625-639.

Michelacci, C., & Silva, O. (2007). Why so many local entrepreneurs?. The Review of Economics and Statistics, 89(4), 615-633.

Partridge, M.D., Bollman, R.D., Olfert, M.R. and Alasi, A. (2007), Riding the wave of Urban Growth in the Countryside: Spread, Backwash or Stagnation? Land Economics, 83(2), 128-152.

Pettersson, L., Sjölander, P., & Widell, L. M. (2010). Firm formation in rural and urban regions explained by demographical structure. Paper presented at the 50th European Regional Science Association (ERSA) Conference, Jönköping, Sweden, 19-23 August 2010.

Sandow, E. (2008). Commuting behaviour in sparsely populated areas: evidence from northern Sweden. Journal of Transport Geography, 16(1), 14-27.

Stam, E. (2009). Entrepreneurship, evolution and geography. Papers in Evolutionary Economic Geography, 9(13), 1-23.

Statistics Netherlands (2016). Banen werknemers en afstand woon-werk; woon- en werkregio's. Retrieved through Statline (http://statline.cbs.nl).

Sternberg, R. (2011). Regional determinants of entrepreneurial activities - theories and empirical evidence. In M. Fritsch (Ed.), Handbook of Research and Entrepreneurship and Regional Development (pp. 33-57). Cheltenham/Northampton: Edward Elgar Publishing.

Tobler W. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(2), 234-240.

Van Oort, F.G. & Atzema, O. (2004). On the conceptualization of agglomeration economies: the case of new firm formation in the Dutch ICT sector. The Annals of Regional Science, 38, 263-290.

Van Oort, F.G. & Stam, F.C. (2006). Agglomeration Economies and Entrepreneurship in the ICT Industry (No. ERS-2006-016-0RG). ERIM report series research in management Erasmus Research Institute of Management. Erasmus Research Institute of Management (ERIM). Retrieved from http://hdl.handle.net/1765/7639

Van Stel, A. and K. Suddle (2008). The impact of new firm formation on regional development in the Netherlands. Small Business Economics, 30(1), 30-47.

Verhetsel, A., Van Hecke, E., Thomas, I., Beelen, M., Halleux, J.M., Lambotte, J.M., Rixhon, G. & Mérenne-Schoumaker, B. (2009). Pendel in België. De woon-werkverplaatsingen. De woon-schoolverplaatsingen. Brussel: Statistics Belgium, FOD Economie.

Verheul, I., Wennekers, S., Audretsch, D., & Thurik, R. (2002). An eclectic theory of entrepreneurship: policies, institutions and culture. In D. Audretsch, R. Thurik, I. Verheul & S. Wennekers (Eds.) Entrepreneurship: Determinants and policy in a European-US comparison (pp. 11-81). Springer: US.

Zipf, G. (1949). Human Behavior and the Principle of Least Effort. Cambridge MA: Addison-Wesley.

Appendix I

Stata do-file for spatial lag calculation for the Netherlands

```
*set directory
cd "C:\Users\mathi\OneDrive\Documenten\Master thesis\Stata\Relative geography (1 JULI)\NL"

*import data from shapefile
shp2dta using Nederland2013, database(NL2013) coordinates(NL2013xy) genid(id) replace

use NL2013

*destring coordinate variables
quietly destring Longitude, replace
quietly destring Latitude, replace

*enter popdens in data editor***

******Normalized matrices****
*Continuity matrix
spmat contiguity cNLnorm using NL2013xy, id(id) normalize(row) replace
spmat summarize cNLnorm
spmat summarize cNLnorm, links

*deal with islands: export neighborlist and adjust this so that islands neighbour to the municipalities from which boats leave
spmat export cNLnorm using NBcont, nlist replace

spmat import cNLnorm using NBcont, nlist normalize(row) replace
spmat summarize cNLnorm
spmat summarize cNLnorm, links

*Inverse distance matrix - without cut-off; not very sensible as it implies that all municipalities influence each other
spmat idistance iNLnorm Longitude Latitude, id(id) dfunction(dhaversine) normalize(row) replace
spmat summarize iNLnorm
spmat summarize iNLnorm, links

*Inverse distance matrix 30km cutoff
spmat idistance icut30NLnorm Longitude Latitude, id(id) dfunction(dhaversine) vtruncate(1/30) normalize(row) replace
spmat summarize icut30NLnorm
spmat summarize icut30NLnorm, links

*Inverse distance matrix 50km cutoff
spmat idistance icut50NLnorm Longitude Latitude, id(id) dfunction(dhaversine) vtruncate(1/50) normalize(row) replace
spmat summarize icut50NLnorm
spmat summarize icut50NLnorm, links
```

```
*Inverse distance matrix 70km cutoff
spmat idistance icut70NLnorm Longitude Latitude, id(id) dfunction(dhaversine) vtruncate(1/70) normalize(row) replace
spmat summarize icut70NLnorm
spmat summarize icut70NLnorm, links

*Inverse distance matrix 100km cutoff
spmat idistance icut100NLnorm Longitude Latitude, id(id) dfunction(dhaversine) vtruncate(1/100) normalize(row) replace
spmat summarize icut100NLnorm
spmat summarize icut100NLnorm, links

*Inverse distance matrix 120km cutoff
spmat idistance icut120NLnorm Longitude Latitude, id(id) dfunction(dhaversine) vtruncate(1/120) normalize(row) replace
spmat summarize icut120NLnorm
spmat summarize icut120NLnorm, links

*Inverse distance matrix 250km cutoff
spmat idistance icut250NLnorm Longitude Latitude, id(id) dfunction(dhaversine) vtruncate(1/250) normalize(row) replace
spmat summarize icut250NLnorm
spmat summarize icut250NLnorm, links

* writing matrix to a text file
spmat export cNLnorm using cNL, replace
spmat export iNLnorm using idNL, replace
spmat export icut30NLnorm using idNL30, replace
spmat export icut50NLnorm using idNL50, replace
spmat export icut70NLnorm using idNL70, replace
spmat export icut100NLnorm using idNL100, replace
spmat export icut120NLnorm using idNL120, replace
spmat export icut250NLnorm using idNL250, replace

*Create spatial lag of population density variable with continuity matrix
spmat lag contlag cNLnorm Popdens2013

*Create spatial lag of population density variable with inversedistance matrix
spmat lag idlag iNLnorm Popdens2013

*Create spatial lag of population density variable with inversedistance matrix 30 km cutoff
spmat lag id30lag icut30NLnorm Popdens2013

*Create spatial lag of population density variable with inversedistance matrix 50 km cutoff
spmat lag id50lag icut50NLnorm Popdens2013

*Create spatial lag of population density variable with inversedistance matrix 70 km cutoff
spmat lag id70lag icut70NLnorm Popdens2013

*Create spatial lag of population density variable with inversedistance matrix 100 km cutoff
spmat lag id100lag icut100NLnorm Popdens2013
```

```
*Create spatial lag of population density variable with inversedistance matrix 120 km cutoff
spmat lag id120lag icut120NLnorm Popdens2013

*Create spatial lag of population density variable with inversedistance matrix 250 km cutoff
spmat lag id250lag icut250NLnorm Popdens2013

*save file
save "C:\Users\mathi\OneDrive\Documenten\Master thesis\Stata\Stata GIS NL\31 mei NL\NL2013lags.dta", replace
```

Appendix II

Stata do-file for the Maximum Likelihood Estimation

```
cd "C:\Users\mathi\OneDrive\Documenten\Master thesis\Stata\Testen\definitief

program define normal
version 1.0
args lnf mu sigma
quietly replace `lnf'=ln(normd(($ML_y1-`mu')/`sigma'))-ln(`sigma')
end

**** APART FOR THE THREE COUNTRIES (table 3) with natural log, distance squared termen and centered to control for multicollinearity*****

****NETHERLANDS
clear
use "Datasetcompleet NL y x"

**model 1
ml model lf normal (RateLM0813=c_LGpd AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleDEF2.doc, replace ctitle("NL 1")
estat ic

**model 2
ml model lf normal (RateLM0813=c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleDEF2.doc, append ctitle("NL 2")
estat ic

*corelation matrix
estat vce, correlation

*vif
collin c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813

****BELGIUM
clear
use "Datasetcompleet BE y x"
```

```
**model 1
ml model lf normal (RateLM0813=c_LGpd AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleDEF.doc, append ctitle("BE 1")
estat ic

**model 2
ml model lf normal (RateLM0813=c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleDEF.doc, append ctitle("BE 2")
estat ic

*corelation matrix
estat vce, correlation
*vif
collin c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 lnINCOME Unemp0813 Service0813 Public0813

*without wallonie
drop if Wallonie == 1

** model 1
ml model lf normal (RateLM0813=c_LGpd AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleDEF2robust.doc, replace ctitle("BE 1")
estat ic

**model 2
ml model lf normal (RateLM0813=c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleDEF2robust.doc, append ctitle("BE 2")
estat ic

*corelation matrix
estat vce, correlation
*vif
collin RateLM0813 c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813
```

```
*with wallonie dummy
clear
use "Datasetcompleet BE y x"

**model 1
ml model lf normal (RateLM0813=c_LGpd AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 Wallonie )(RateLM0813=)
ml maximize
outreg2 using mleDEF2bedum.doc, replace ctitle("bE 1")
estat ic

**model 2
ml model lf normal (RateLM0813=c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 Wallonie )(RateLM0813=)
ml maximize
outreg2 using mleDEF2bedum.doc, append ctitle("bE 2")
estat ic


****SWEDEN
clear
use "Datasetcompleet SE y x"

**model 1
ml model lf normal (RateLM0813=c_LGpd AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleDEF2.doc, append ctitle("SE 1")
estat ic

**model 2
ml model lf normal (RateLM0813=c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleDEF2.doc, append ctitle("SE 2")
estat ic

*corelation matrix
estat vce, correlation
*vif
collin RateLM0813 c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813
```

```
**without north of Sweden
drop if NORD_SE == 1

**model 1
ml model lf normal (RateLM0813=c_LGpd AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleSE.doc, replace ctitle("SE 1")
estat ic

**model 2
ml model lf normal (RateLM0813=c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 )(RateLM0813=)
ml maximize
outreg2 using mleSE.doc, append ctitle("SE 2")
estat ic

**with north of Sweden dummy
clear
use "Datasetcompleet SE y x"

**model 1
ml model lf normal (RateLM0813=c_LGpd AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 NORD_SE)(RateLM0813=)
ml maximize
outreg2 using mleDEF2.doc, append ctitle("SE 1")
estat ic

**model 2
ml model lf normal (RateLM0813=c_LGpd id50lag c_distnear c_distnear_square c_dist50 c_dist50_square c_dist100 c_dist100_square c_dist250 c_dist250_square AvShPop0813_under15 AvShPop0813_15till25 AvShPop0813_25till35 AvShPop0813_50till65 AvShPop0813_65plus EduHigh0813 Immi0813 Unemp0813 Service0813 Public0813 NORD_SE)(RateLM0813=)
ml maximize
outreg2 using mleDEF2.doc, append ctitle("SE 2")
estat ic
```