Mapping the Spatial Pattern of the Uncertain Data: A comparison of Global Non-response Rate (GNR) between Metropolitan and Non- Metropolitan Area

Scott Bell and Ting Wei

Department of Geography and Planning, University of Saskatchewan, Saskatoon, SK

Abstract

Statistics Canada has recently invoked substantial changes in the way they collect census data regarding the Canadian populace. Canada has shifted from a legally-enforced long form with a 20% sample to a voluntary National Household Survey (NHS) with a 30+% sample (Canada, 2011a). This change raises the concern regarding the spatial variability of uncertainty in NHS data across Canada. Understanding the distribution of data uncertainty is essential for researchers to consider, particularly if they are considering using such data for research. Furthermore, this will provide insights for improving the sampling methods to better represent the total population. Both descriptive statistics and spatial autocorrelation analysis are used to examine the Global Non-response Rate (GNR, which is used a data quality indicator for NHS data) variation at different urbanization levels within each province and for Canada as a whole. Results indicate that GNR is lower in metropolitan areas than non-metropolitan areas. Local spatial autocorrelation analysis indicates that low-low GNR clusters tend to appear in metropolitan areas, while high-high GNR clusters are more likely to be in non- metropolitan areas.

Background and Relevance

Understanding data uncertainty is central to communicating patterns present in populations represented by sampled data. For decision makers (planners, policy makers, elected, and non-elected officials) the need to understand underlying demographic patterns is essential to making informed decisions. Understanding, or at least having access to, the range of possible outcomes of a sampled dataset plays an important role in the likelihood of invoking a decision, the voracity with which a spatial pattern is defended, or even the likelihood of it being used as part of the decision making process. Decision making in Canadian cities is often based on emerging patterns of population change summarized by census data. Statistics Canada has recently invoked substantial changes in the way they collect census data regarding the Canadian populace. Canada has shifted from a legally-enforced long form with a 20% sample to a voluntary National Household Survey with a 30+% sample (Canada, 2011a). This change demands that we examine the availability as well as variability of uncertainty of the collected data (Kardos, Benwell, & Moore, 2005). It is important to note that this research is not about the uncertainty of spatial data, but the spatial nature of data uncertainty. The focus of this research is to explore the variability of uncertainty for metropolitan and non- metropolitan area in Canada and individual provinces.

Methods and Data

The key variable used in this research is the Global Non-response Rate (GNR) of the National Household Survey 2011 use Census SubDivisions (CSD) as the unit of analysis. We have categorize the study area into three types of geography: Census Metropolitan Area (CMA), Peri-Urban Area, and Rural area. CMA boundaries are used to define the extent of the urban space, CSDs that touch a CMA boundary are defined as Peri-Urban space, and all other CSDs are rural. GNR combines complete non-response (household) and partial non-response (question) into a single rate, and is used as an indicator of data quality (Canada, 2011b). Smaller values of GNR indicate lower risk of inaccuracy. According to NHS user guide (Canada, 2011a), products of any geographic areas with GNR greater or equal to 50% is not released due to the high level of error which exceed an acceptable threshold. As a result, Census SubDivisions without GNR are excluded from the analysis and mapping processes. Descriptive statistics are used to summarize the basic features of GNR among the three study geographies in Canada and individual provinces. GNR values have been joined to CSDs for spatial autocorrelation analysis. Local Indicators of Spatial Association (LISA) based on local Moran statistic (Anselin, 1995) is used to examine if clustering of similar values exists in the spatial arrangement of GNR.

Results

Descriptive statistics of GNR

In early research, we mapped GNR at national and provincial level (Bell, Jones, & Wei, 2013). Both Peri-Urban and Rural spaces are different from urban spaces in terms of GNR value and availability. These maps and analysis revealed the impact of population density on data variability and suggest this is a pressing problem for users of this data. This visual examination based on choropleth maps is consistent with mean comparison results of GNR at three urbanization levels in Canada and individual provinces. Table 1 provides descriptive statistics of GNR for Census Metropolitan, Peri-Urban and Rural geographies. Generally, GNR fluctuates around 30% for provinces. In remote areas, all the households were invited to participate the NHS 2011 survey (Canada, 2011a). Consequently, it is not surprising that Northwest, Nunavut, and Yukon territories have relatively lower GNR. In terms of urbanization, Peri-Urban and Rural geographies have higher GNR than Census Metropolitan geographies; this pattern is true for most provinces (figure 1). In other words, densely populated metropolitan areas have higher response rates than more sparsely populated non-metropolitan areas. In addition, it is notable that Peri-Urban places seems to have the higher GNR mean values compared to Rural geographies for most of the provinces.

Table 1. Descriptive statistics of GNR by province and three levels of urbanization

Province	CMA		Peri-Urban		Rural		total	
	N	Mean (sd)	N	Mean (sd)	N	Mean (sd)	N	Mean (sd)
AB	61	29.06 (11.53)	28	36.28 (11.80)	204	30.29 (16.36)	293	30.61 (15.16)
BC	164	26.67 (10.60)	46	34.45 (12.05)	227	28.13 (14.09)	437	28.25 (12.84)
MB	17	29.76 (8.75)	23	40.65 (7.01)	150	28.25 (15.03)	190	29.88 (14.37)
NB	63	30.06 (9.93)	42	40.43 (7.60)	86	34.86 (10.23)	191	34.50 (10.29)
NL	22	35.49 (9.32)	5	31.98 (15.44)	214	31.78 (12.25)	241	32.12 (12.08)
NT	1	14.7	2	14.55 (3.75)	31	16.67 (7.41)	34	16.49 (7.12)
NS	19	23.20 (9.73)	10	34.34 (6.92)	47	32.01 (12.14)	76	30.11 (11.64)
NU	-	-	-	-	21	21.21 (9.08)	21	21.21 (9.08)
ON	139	29.08 (8.00)	104	37.69 (8.13)	186	30.65 (13.56)	429	31.85 (11.29)
PEI	15	35.77 (9.68)	10	34.64 (15.78)	52	39.13 (9.79)	77	37.89 (10.70)
QC	215	26.39 (9.84)	227	36.07 (8.91)	536	34.25 (10.48)	978	32.94 (10.60)
SK	38	35.13 (11.46)	27	37.19 (11.64)	391	33.02 (13.10)	456	33.45 (12.92)
YK	2	37.35 (17.04)	-	-	13	24.80 (10.49)	15	26.47 (11.60)
Canada	756	28.37 (10.25)	524	36.69 (9.60)	2158	31.69 (13.15)	3438	31.72 (13.32)

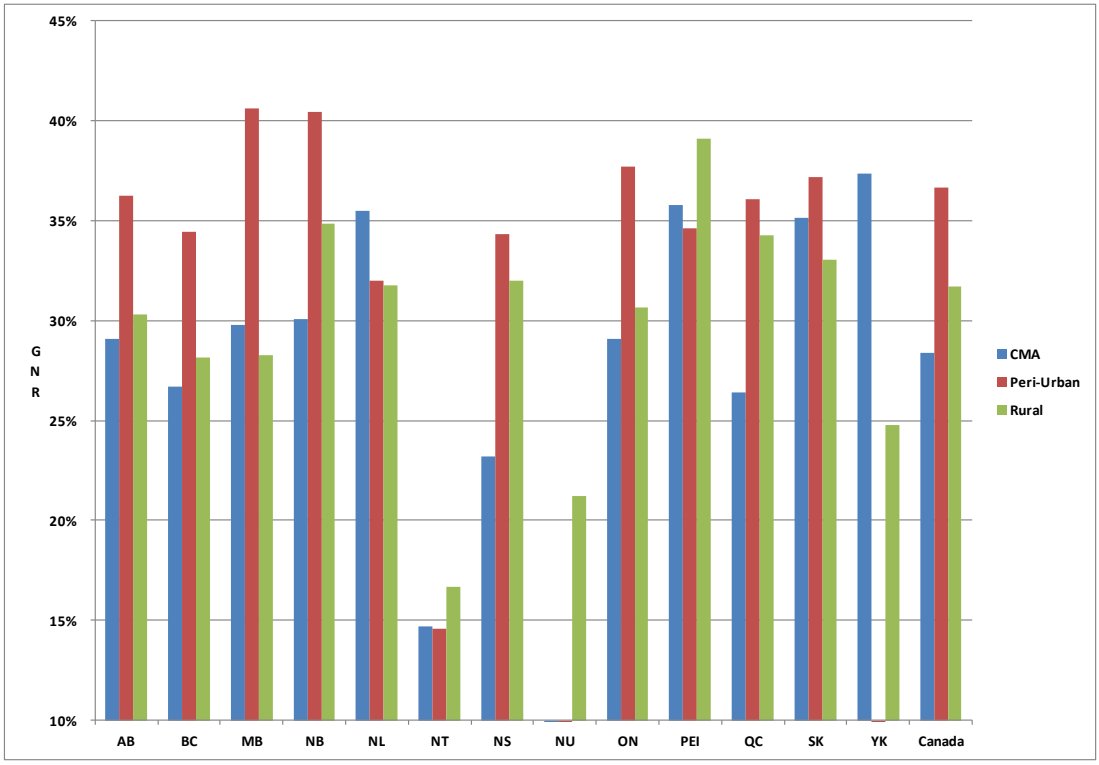
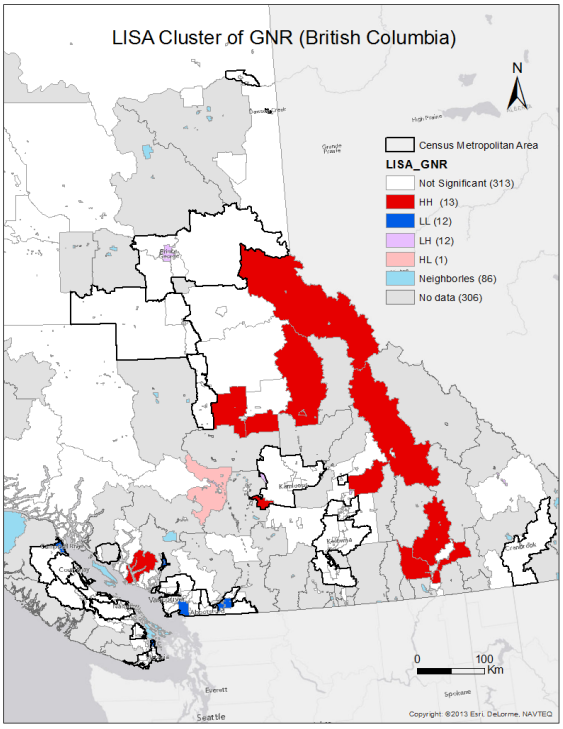


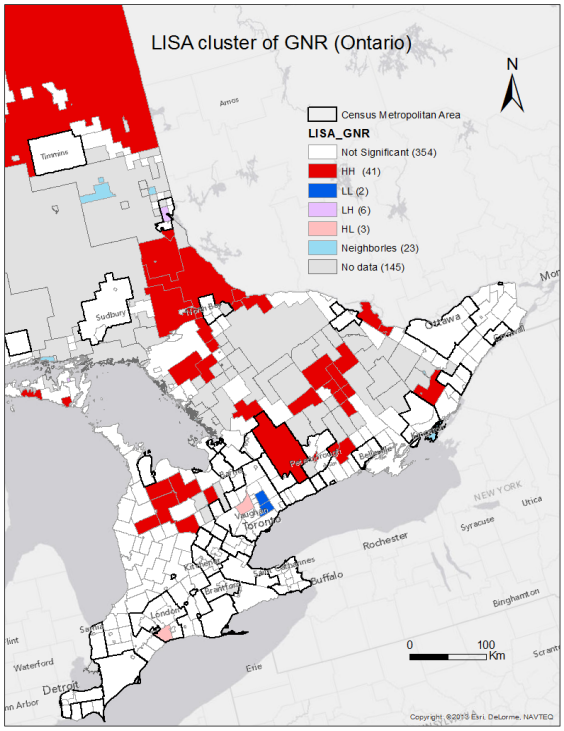
Figure 1. Average GNR by province and three levels of urbanization

Spatial Arrangement of GNR

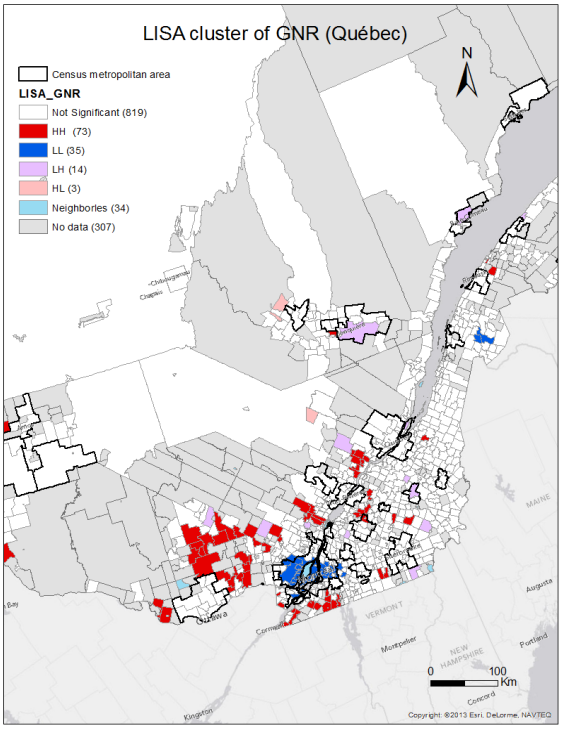
Figure 2a, 2b and 2c are LISA cluster maps from the spatial autocorrelation analysis for British Columbia, Ontario, and Québec. These three provinces were selected because they are the most populous in Canada. In addition to CSDs without GNR values (shown as "No data"), Census SubDivisons are classified into five categories based on the results: Not significant (Census SubDivisons that are not significant at the level of 0.05); HH (high GNR values surrounded by high GNR values); LL (low GNR values surrounded by low GNR values); LH (low GNR values surrounded by high GNR values); HL (high GNR values surrounded by low GNR values) and Neighbourless (areas that are surrounded by Census SubDivisions without GNR). Overall, Low-Low clusters appear in the largest metropolitan areas, including Vancouver, Toronto and Montreal, while High-High cluster appear in non-metropolitan areas. Low-Low clusters are particularly apparent for the areas near Montreal in Québec (see figure 2c). Furthermore, it is worth mentioning that a large number of High-High clusters are surrounded by areas with unavailable GNR (actual value is greater or equal to 50%), inflating the importance of the remaining neighbours in the calculation of this local statistic. This was a concern of data qualify of NHS 2011 in rural area due to low response rate in non-metropolitan area.



(a)



(b)



(c)

Figure 2. LISA Cluster Map of British Columbia (a), Ontario (b) and Québec (c) (GNR at Census SubDivision level)

Conclusions

In this preliminary analysis of GNR, both descriptive statistics and spatial autocorrelation analysis are used to evaluate the statistical and spatial variability of GNR at three geographies in Canada and for individual provinces. Results show that average GNR for metropolitan geographies is lower than non-metropolitan. Surprisingly, the average GNR of Peri-Urban areas is even higher than rural. In terms of the spatial arrangement of GNR, Low-Low clusters are more likely to appear in metropolitan areas while High-High cluster tend to be in non-metropolitan. Results of this research reveal the variation of GNR between metropolitan and non-metropolitan areas, which will lead to future exploration of factors causing of such difference. Future work could include comparing the social-demographic differences among Census Subdivisions with varied GNR values. Research suggests that socio-demographic status influences census response rate (Vigdor, 2004). Results of this comparison will provide insights for predicting the Global Non- Response rate. Also, characteristics of Census SubDivisons with high GNR can be used for improving the sampling methods in order to better targeting population who are unlikely to participate.

References

Anselin, L. (1995). Local indicators of spatial association—LISA. *Geographical analysis, 27*(2), 93-115.

Bell, S., Jones, M., & Wei, T. (2013). *Communicating with uncertain data: A search for spatial and geographic patterns* Paper presented at the COSIT 2013 Workshop on Visually-Supported Reasoning with Uncertainty, Scarborough, North Yorkshire, UK.

Canada, S. (2011a). National Household Survey User Guide, 2011. Retrieved October 24th, 2013, from http://www12.statcan.gc.ca/nhs-enm/2011/ref/nhs-enm_guide/99-001-x2011001-eng.pdf.

Canada, S. (2011b). NHS Profile, 2011 – About the data. Retrieved October 24th, 2013, from http://www12.statcan.gc.ca/nhs-enm/2011/dp-pd/prof/help-aide/aboutdata-aproposdonnees.cfm?Lang=E.

Kardos, J., Benwell, G., & Moore, A. (2005). The visualisation of uncertainty for spatially referenced census data using hierarchical tessellations. *Transactions in GIS, 9*(1), 19-34.

Vigdor, J. L. (2004). Community composition and collective action: Analyzing initial mail response to the 2000 census. *Review of Economics and Statistics, 86*(1), 303-312.