Using Principal Component Analysis to Identify Priority Neighbourhoods for Health Services Delivery by Ranking Socioeconomic Status

Christine Elizabeth Friesen[1], Patrick Seliske[2], Andrew Papadopoulos[3]

1. University of Guelph, Guelph, Ontario, Canada
2. Wellington-Dufferin-Guelph Public Health, Guelph, Ontario, Canada
3. University of Guelph, Guelph, Ontario, Canada

Abstract

Objectives. Socioeconomic status (SES) is a comprehensive indicator of health status and is useful in area-level health research and informing public health resource allocation. Principal component analysis (PCA) is a useful tool for developing SES indices to identify area-level disparities in SES within communities. While SES research in Canada has relied on census data, the voluntary nature of the 2011 National Household Survey challenges the validity of its data, especially income variables. This study sought to determine the appropriateness of replacing census income information with tax filer data in neighbourhood SES index development. **Methods.** Census and taxfiler data for Guelph, Ontario were retrieved for the years 2005, 2006, and 2011. Data were extracted for eleven income and non-income SES variables. PCA was employed to identify significant principal components from each dataset and weights of each contributing variable. Variable-specific factor scores were applied to standardized census and taxfiler data values to produce SES scores. **Results.** The substitution of taxfiler income variables for census income variables yielded SES score distributions and neighbourhood SES classifications that were similar to SES scores calculated using entirely census variables. Combining taxfiler income variables with census non-income variables also produced clearer SES level distinctions. Internal validation procedures indicated that utilizing multiple principal components produced clearer SES level distinctions than using only the first principal component. **Conclusion.** Identifying socioeconomic disparities between neighbourhoods is an important step in assessing the level of disadvantage of communities. The ability to replace census income information with taxfiler data to develop SES indices expands the versatility of public health research and planning in Canada, as more data sources can be explored. The apparent usefulness of PCA also contributes to the improvement of SES measurement and calculation methods, and the freedom to input area-specific data allows the present method to be adapted to other locales.

Keywords: **area-level socioeconomic status, principal component analysis, priority neighbourhood, census, national household survey**

Correspondence: cfries02@uoguelph.ca

DOI: 10.5210/ojphi.v8i2.6733

INTRODUCTION

Determinants of Health

Income has been used in public health research and practice for many years as an indicator of health status and predictor of health outcomes. Low income has been associated with an increased risk of developing chronic conditions like arthritis and diabetes, living with a disability, and experiencing mental health issues [1,2]. Pre-existing medical conditions may perpetuate a vicious cycle by restricting exposure to employment opportunities [3]. In Canada, being low-income, as opposed to middle- or high-income, is associated with increased use of health care resources [1], and restricted access to privately-funded medical procedures, dental coverage, screening services, educational resources, affordable housing, and safe working environments [2,3]. At the population level, a large income disparity between wealthy and poor individuals in a community has been linked to increased rates of disease across the population and high costs to the medical system [1,3].

While income remains a very important determinant of health, other factors like education, employment, and family structure can significantly affect the health of individuals and the community as a whole. Socioeconomic status (SES) has been used as a predictor of health outcomes and is more comprehensive than income alone. SES encompasses the conditions experienced by individuals and communities created by complex interactions between income, employment, occupation, education level, and family dynamics [4-11].

The relationship between education and income has been shown in several studies [12-15]. For instance, low motivation to pursue education can be the result of family background, poor family standard of living, parental structure, and low educational aspirations by parents [12]. Those who have not completed a high school education may not achieve the level of verbal skills nor be exposed to employment opportunities that are associated with the attainment of high-paying jobs [14,15]. Additionally, attaining a high school diploma is becoming more recognized by Canadians as a requirement for many training programs as well as the common prerequisite for joining the labour force [13]. As a result, those who do not complete high school may be less able to afford safe housing and healthy food, leaving them at higher risk for negative health outcomes and criminal behaviours [3,13-15].

Job loss is a stressful event that affects self-esteem and can lead to harmful coping mechanisms, such as substance abuse and engaging in criminal activities [3]. Unemployed individuals suffer high mortality rates and more severe health problems than those who are employed [16]. When seeking a new job, these individuals may be more likely to accept lower-paying employment in more dangerous conditions [3].

Lone parent status is another important social determinant of health that influences SES. Families led by a single parent are more likely to be classified as low income and are among the most impoverished priority populations [1,2]. When the parent does not have a well-paying occupation, the family is at an increased risk for poor health outcomes resulting from lack of access to health and educational services, good housing, safe working environments, and healthy food [1]. Compared to single fathers, single mothers are at even higher risk for poor health for themselves and their families, due to pay inequalities between men and women, workplace discrimination, increased psychosocial pressures, and costly childcare [3,7].

Socioeconomic Status and Principal Component Analysis

The objective definition of and research into SES is relatively new due to its complex nature. Historically, social status was measured on simple scales that allocated an equal weight to individual-level factors such as occupation type, occupation of friends, income, and education level [8]. However, it has become apparent that social conditions outside of an individual's direct control can influence one's health [5]. Socioeconomic indicators of health cluster at the neighbourhood level, which contribute to the understanding of health inequities within communities [7]. Principal component analysis (PCA) is a statistical technique that has been used to develop area-level SES indices that are often mapped using geographic information systems to produce clear visual boundaries of SES differentials [4,6,10,11,17]. This information informs public health resource allocation, service delivery, and program dissemination as it provides a more comprehensive understanding of communities' levels of disadvantage in relation to one another.

Relevant SES variables may be inputted into a PCA-capable program to extract multiple underlying dimensions based on the variation produced by these correlated variables. Common statistical assumptions of normality and homoscedasticity do not apply to PCA, which eliminates the need for data transformations that often result in a loss of original information [18]. PCA outputs a list of principal components (PCs) that are independent orthogonal linear combinations of the variables and are listed in decreasing order of proportion of explained variance.

The first PC produced when utilizing SES indicator variables (such as income and unemployment) has often been considered the only dimension pertaining to SES, and therefore only the variable loadings pertaining to the first PC have been used for SES calculations previously [5,7,9-11,19]. Other researchers, however, have used variable loadings from any components that each represents a sufficient proportion of the overall

variation [4,6]. The literature generally supports the notion that the first PC represents the economic system aspect of SES [4-7,9-11,19], while subsequent components may represent other dimensions of SES, such as the social system and marginalization [6], depending on which variables contribute highly to that principal component [4].

The use of PCA has been important to the development of indices because it assigns different weights to each variable, as opposed to arbitrarily weighing each variable equally [2]. While it is simpler to assign equal weights to variables in an index, modern understanding of SES requires the exploration of complex relationships between variables that historically simple methods do not support. Furthermore, PCA can provide insight into which variables have greater influence on the dimension(s) of SES [4,6] when using area-specific data to inform public health policy, interventions, and resource allocation according to the area's unique needs.

Changes to the Canadian Census Affecting SES Index Development

SES research by Wellington-Dufferin-Guelph Public Health [2] has been performed in the past to identify priority neighbourhoods in their service area [2]. This research was reliant on accurate census information pertaining to income and non-income variables contributing to SES. Due to privacy-related concerns and decreasing response rates, the Canadian Order in Council decided in June 2010 that the mandatory 2011 Canadian Census would only collect demographic information pertaining to family structure, spoken language, and farm management practices [20]. A second voluntary National Household Survey (NHS) was distributed that resembled the 2006 long-form Census [21]. The validity of data collected by the 2011 NHS, namely income fields, is questionable due to an increased risk of non-response bias stemming from the voluntary nature of the NHS. Taxfiler data may prove to be a viable alternative to current NHS data in the calculation of SES at the neighbourhood level. The federal government acquires taxfiler information annually and the completion of personal tax returns is mandatory by individuals in the labour force. Taxfiler data in Canada are more precise than census estimates in terms of dollar and cent amounts, and must be completed truthfully to avoid fines and penalties, which is further cross-referenced with submitted employer records.

The current report presents a method of developing neighbourhood-level SES indices using PCA with income and non-income variables indicative of SES. This study assesses the effectiveness of using taxfiler data as an alternative data source to replace current Canadian census income variables typically used in SES calculations. The present study also seeks to descriptively validate the use of multiple PCs in the calculation of SES and support the inclusion of non-income variables. This method can be tailored to other locales by selecting variables appropriate for the local demographics and the results may be utilized to guide the allocation of resources that support the health of priority populations.

METHODS

Data Sources

Census and taxfiler data for the predominantly urban Census Metropolitan Area (CMA) of the city of Guelph were obtained through WDGPH's membership as part of the Community Data Consortium, which permits access to the Canadian Council on Social Development's Community Data Program (CDP) [22]. Data for the current project were obtained at the census tract (CT) level, which geographically divides a CMA with a core population of 50,000 or more into smaller areas containing 2,500 to 8,000 persons. CTs attempt to contain individuals that are generally homogenous in terms of living conditions and socioeconomic characteristics [23,24]. Census Profiles for the years 2006 and 2011 were obtained in addition to the 2011 NHS Profile. Taxfiler Family Data tables were obtained for the years 2005 and 2011. Taxfiler data from 2005 were considered comparable to the 2006 Census Profile in terms of income fields, since the 2006 Census income measurements were based on individuals' assessment of their 2005 incomes. Appendix A presents a description of each dataset. Ethics approval was obtained through the University's Research Ethics Board.

Data Processing

Census and NHS data were extracted from the original CDP files using Beyond 20/20 (Beyond 20/20, 2015). Taxfiler datasets were provided in Microsoft Excel (Microsoft, 2013) format. All datasets were further processed within Microsoft Excel, where data were restricted to the CTs within the borders of the Guelph CMA and fields pertinent to SES Indicator Variables were retained (Table 1). Twenty-one CTs based on 2001 Census geographies were retained for the 2005 taxfiler dataset. Twenty-seven CTs were retained for 2006 Census and all 2011 datasets. In order to compare 2005 taxfiler income variables with those from the 2006 Census, the additional CTs in 2006 were combined and averaged according to the previous 2001 Census CT geographies.

Table 1. Income and non-income socioeconomic indicator variables derived from five datasets for the years 2005, 2006, and 2011 for the census metropolitan area of Guelph.

Income Indicator	Description	Source Datasets
Median Family Income ($)	Calculated by Statistics Canada	Taxfiler Family Data, 2005
		Taxfiler Family Data, 2011
		Census Profile, 2006
Median Single Person Income ($)	Calculated by Statistics Canada	Taxfiler Family Data, 2005
		Taxfiler Family Data, 2011
		Census Profile, 2006
Low Income Families, After-tax (%)	$\left(\frac{\text{LI Couple Fam.} + \text{LI Lone Parent Fam.}}{\text{Couple Fam.} + \text{Lone Parent Fam.}}\right) \times 100\%$	Taxfiler Family Data, 2005
	$\left(\frac{\text{LI Couple Fam.} + \text{LI Lone Parent Fam.}}{\text{Couple Fam.} + \text{Lone Parent Fam.}}\right) \times 100\%$	Taxfiler Family Data, 2011
	Calculated by Statistics Canada	Census Profile, 2006
Low Income Unattached, After-tax (%)	$\left(\frac{\text{LI Single Person}}{\text{Single Person}}\right) \times 100\%$	Taxfiler Family Data, 2005
	$\left(\frac{\text{LI Single Person}}{\text{Single Person}}\right) \times 100\%$	Taxfiler Family Data, 2011
	Calculated by Statistics Canada	Census Profile, 2006
Non-Income Indicator	**Description**	**Source Datasets**
Lone Parent Families (%)	$\left(\frac{\text{Lone Parent Families}}{\text{Total Families}}\right) \times 100\%$	Census Profile, 2006
		Census Profile, 2011
Single Mothers (%)	$\left(\frac{\text{Female} - \text{Led Lone Parent Fam.}}{\text{Total Families}}\right) \times 100\%$	Census Profile, 2006
		Census Profile, 2011
Unemployment Rate, 15 years and over (%)	Calculated by Statistics Canada	Census Profile, 2006
		NHS Profile, 2011
Low Education, 15 years and over (%)	$\left(\frac{\text{LE Over 15 (20\% of Sample)}}{\text{Total Population (20\% of Sample)}}\right) \times 100\%$	Census Profile, 2006
		NHS Profile, 2011
Average Home Value ($)	Calculated by Statistics Canada	Census Profile, 2006
		NHS Profile, 2011
Average Monthly Rent ($)	Calculated by Statistics Canada	Census Profile, 2006
		NHS Profile, 2011
Managerial Occupation (%)	$\left(\frac{\text{Managerial Position}}{\text{Total Work Force}}\right) \times 100\%$	Census Profile, 2006
		NHS Profile, 2011

Variable Selection

Eleven variables were retained and calculated from the five source datasets (Table 1). The following nine variables were selected based on previous research by WDGPH [2]: median family income; median single person income; proportion of low income families; proportion of low income unattached; proportion of lone parents; unemployment rate of those aged 15 or older; proportion of those aged 15 or older with a low education level; average home value; and average monthly rent. Households were considered 'low income' if they fell into an income threshold in which more than 20% of their income was expended on food, clothing, and shelter [25]. Previous WDGPH research restricted 'low education' to individuals with 'less than grade 9 education'; however, evidence from the literature suggests that less than a high school education is associated with adverse health and economic outcomes [3,12-15]. The present study therefore utilized the census field 'No certificate, degree or diploma' to represent individuals with relatively low education. Two additional non-income variables were included in the present analysis: proportion of single mothers [3], and proportion of individuals in a managerial occupation [6,7,9].

Statistical Analyses

Descriptive Statistics

Means, ranges, and standard deviations for each of the 11 indicator variables were obtained using STATA 13 (StataCorp LP, 2013). The results are presented in Table 2 stratified by data source.

Taxfiler Income Variable Validation

The first PCA was performed in STATA 13 using the 11 indicator variables from the 2006 Census Profile. Data were automatically standardized on the correlation matrix using the 'pca' function in STATA 13. The resulting PCs were selected for further analysis if they first met Kaiser's Criterion, where PCs with an eigenvector greater than 1.0 should be retained [6,11]. PCs that met Kaiser's Criterion were excluded if they represented less than 10% of the variance from the original variables [4]. For the purposes of later calculations, each selected PC was weighted according to its proportion of the sum of the variance represented by the selected PCs:

$$\mathbf{PC_w} = \frac{\mathbf{PC_{Proportion\ of\ Total\ Variance}}}{\mathbf{Proportion\ of\ Total\ Variance\ Represented\ by\ Selected\ PCs}} \quad \mathbf{(1)}$$

In order to interpret influential variables amongst each retained PC, un-rotated eigenvector correlations, or variable loadings, were examined. Signs indicated the direction of a variable's influence on the underlying dimension explained by the PC relative to the

influence of the other variables. Hair *et al.* [26] suggested that |0.4| should be the lowest cut-off for relevant variable loadings, while central factors should have eigenvectors of at least |0.6|. However, recent research suggests that data realistically produces minimum relevant variable loadings closer to |0.25| with the central factor(s) having an eigenvector of around |0.4| [27]. The present method considered the suggestions of Raubenheimer [27], which better suited the variable loadings produced by this research. Eigenvectors greater than the absolute average variable loading (Eq. 2) as well as eigenvectors within 0.1 less than the absolute average variable loading were considered influential on each PC:

$$\left| \sqrt{\frac{1}{\#\text{ of Variables}}} \right| \qquad (2)$$

Indicator variable loadings were then multiplied by the corresponding PC weight (PC_w; see Eq. 1) and summed to create variable-specific factor scores to be applied to the CT-specific data values:

$$\text{Factor Score} = \sum (\text{Indicator Variable Loading} \times \text{PC}_\text{w})_{1\ldots\text{j}} \qquad (3)$$

where *Indicator Var. Loading* = eigenvector per selected PC for an indicator variable;
and where PC_w = PC weight for each selected PC.

Data from the 2006 Census Profile were converted to z-scores in STATA 13 to standardize measurement units. The standardized data were multiplied by the variable-specific factor scores and values were summed to create CT-specific SES Scores. SES Scores were then standardized to range from zero to one for easier interpretation [6]:

$$\textit{The SES Scores} \qquad (4)$$

Standardized SES Scores were reversed so that a higher SES Score represented a higher socioeconomic status of a given CT. SES Scores were plotted using a bar graph and divisions in SES Score levels were determined descriptively by visual inspection. In addition to this, cut-offs were used to support the visual aspect by calculating the average numerical increment in SES Score needed to produce a constant increase in SES Score distribution. Differences between subsequent CT SES Scores greater than 0.0476 (for 2005 and 2006 data containing 21 CTs) or 0.037 (for 2011 data containing 27 CTs) were utilized to confirm distinctions in SES Scores made upon visual inspection.

A second PCA was performed and an SES index created by substituting the four income fields from the 2006 Census Profile with four similar fields from 2005 taxfiler data. Validation of the use of taxfiler data involved a descriptive comparison between the two methods of SES Score distributions and CT movements between SES level classifications.

A third SES index was produced according to the above procedure using 2011 Census Profile, 2011 NHS Profile and 2011 taxfiler datasets for the CMA of Guelph.

Internal Validation

Two internal validation procedures were performed. First, SES indices were developed for all three data groups (i.e. 2006 Census only, 2006 Census + 2005 taxfiler, and 2011 Census + 2011 NHS + 2011 taxfiler), using only the first PC. The SES Score distributions of these were compared to the SES Score distributions produced when using all of the components with an eigenvalue greater than 1.0 and a proportion of explained variance greater than 10%. CT movements were also compared between methods.

The second validation procedure involved removing the non-income variables from the SES Score calculation. These SES Score distributions produced were compared to the distributions produced by the 'first PC method' and the 'eigenvalue >1.0 method' to assess the effect of non-income variables on SES level. CT movements between methods were also compared.

RESULTS

1) Descriptive Statistics

The population of the city of Guelph within its CMA increased by 5.6% from 114,943 in the year 2006 to 121,688 in 2011, according to Census Profile data. The 2011 NHS Profile reported a similar population of 120,540. Taxfiler datasets from 2005 and 2011 reported 86,120 and 92,650 tax filers, respectively. Table 2 describes the characteristics of each SES Indicator stratified by data source.

Table 2. Descriptive statistics of the 11 Indicator Variables selected for principal component analysis stratified by source dataset for the census metropolitan area of Guelph for the years 2005, 2006, and 2011.

Indicator	Source Dataset	Median	Mean	*SD*	Range	
					Min.	Max.
Median Family Income ($)	2006 Census Profile	73483.00	73439.20	16947.58	43926.00	108581.00
	2005 Taxfiler	68600.00	68480.95	15218.79	42600.00	97500.00
	2011 Taxfiler	83550.00	80653.70	17144.56	48080.00	113350.00
Median Single Person Income ($)	2006 Census Profile	28189.00	29336.07	7080.29	17231.00	48928.50
	2005 Taxfiler	27000.00	27185.71	3023.95	20100.00	32800.00
	2011 Taxfiler	28960.00	30394.07	5205.80	22430.00	44370.00
Low Income Families, After-tax (%)	2006 Census Profile	6.1	6.4	3.8	1.0	15.9
	2005 Taxfiler	8.4	10.6	6.6	3.8	27.8
	2011 Taxfiler	8.2	10.3	5.8	3.8	26.5
Low Income Unattached, After-tax (%)	2006 Census Profile	25.2	24.6	9.0	7.8	43.5
	2005 Taxfiler	21.7	22.1	5.9	8.8	35.4
	2011 Taxfiler	23.7	24.1	6.5	7.9	35.3
Lone Parent Families (%)	2006 Census Profile	17.4	17.1	5.3	7.2	26.8
	2011 Census Profile	16.3	16.7	5.1	6.0	28.8
Single Mothers (%)	2006 Census Profile	12.6	13.0	4.4	5.5	21.7
	2011 Census Profile	13.0	13.3	4.3	4.3	24.2
Unemployment Rate, 15 years and over (%)	2006 Census Profile	5.1	5.4	1.6	2.3	8.2
	2011 NHS Profile	6.9	6.9	1.7	2.8	10.2
Low Education, 15 years and over (%)	2006 Census Profile	21.6	21.5	6.3	11.4	33.0
	2011 NHS Profile	15.8	17.6	6.0	9.0	31.2
Average Home Value ($)	2006 Census Profile	261031.00	256708.80	48921.65	180472.00	388042.00
	2011 NHS Profile	308219.00	313129.30	52772.46	220035.00	428313.00
Average Monthly Rent ($)	2006 Census Profile	785.00	813.23	78.04	696.00	1014.00
	2011 NHS Profile	844.00	903.41	202.77	640.00	1560.00
Managerial Occupation (%)	2006 Census Profile	7.8	8.6	2.8	3.9	14.8
	2011 NHS Profile	10.1	10.0	2.9	3.9	15.5

SD, standard deviation

2) Taxfiler Income Variable Validation

The results of the first PCA using 2006 Census Profile data are presented in Table 3. The first three retained components cumulatively represent 78.5% of the total variance. The first PC represents 48.5% of the total variance and is highly influenced by family income, proportion of low-income families, proportion of lone parent families and single mothers, average home value, and managerial occupation. To a lesser degree, single person income, unemployment rate, and low education also contribute to PC_1. Median family income could be interpreted as having a negative influence on a community's level of economic deprivation, while the prevalence of low-income families would contribute positively to the community's economic deprivation. Conversely, taking the reciprocal sign may lead to a more intuitive interpretation using economic status, rather than economic deprivation, as the outcome.

The second PC represents an additional 16.4% of the total variance and is strongly influenced by both the proportion of low income unattached individuals and low education; however, these have opposite directional effects on the dimension explained by PC_2. Less influential variables include unemployment rate, single person income, average home value, and managerial occupation.

The third and last retained PC represents 13.7% of the total variance and is mainly influenced by single person income, proportion of single mothers, unemployment rate, and average monthly rent. Family income has a less pronounced influence.

Table 3. Principal component analysis of 11 socioeconomic status indicator variables from 2006 Census Profile data only for the census metropolitan area of Guelph.

Principal Component Eigenvalues					
Component	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**	**PC_w (%)**
PC_1	**5.334090**	**3.535230**	**0.4849**	**0.4849**	**61.74**
PC_2	**1.798860**	**0.292350**	**0.1635**	**0.6484**	**20.82**
PC_3	**1.506500**	**0.488332**	**0.1370**	**0.7854**	**17.44**
C_4	1.018170	0.463475	0.0926	0.8780	—
C_5	0.554698	0.224526	0.0504	0.9284	—
C_6	0.330171	0.156520	0.0300	0.9584	—
C_7	0.173651	0.057958	0.0158	0.9742	—
C_8	0.115693	0.025609	0.0105	0.9847	—
C_9	0.090084	0.034112	0.0082	0.9929	—
C_{10}	0.055972	0.033863	0.0051	0.9980	—
C_{11}	0.022109	—	0.0020	1	—

Indicator Variable Loadings (Eigenvector Correlations)				
Variable	**PC_1**	**PC_2**	**PC_3**	**Factor Score**
Median Family Income ($)	*-**0.3639**	0.0882	*0.2965*	-0.1546
Median Single Person Income ($)	**-*0.2535*	*-0.2387*	**0.5290**	-0.1139
Low Income Families, After-tax (%)	**0.3690**	0.0991	0.0647	0.2597
Low Income Unattached, After-tax (%)	0.1784	**0.6239**	-0.1382	0.2159
Lone Parent Families (%)	**0.4017**	0.1023	0.1536	0.2961
Single Mothers (%)	**0.3307**	0.0151	**0.4462**	0.2851
Unemployment Rate, 15 years and over (%)	*0.2139*	**0.3373**	**0.3641**	0.2658
Low Education, 15 years and over (%)	*0.2076*	**-0.5391**	-0.0063	0.0148
Average Home Value ($)	**-0.372**	*0.2437*	0.1220	-0.1577
Average Monthly Rent ($)	0.0005	0.0060	**0.4885**	0.0868
Managerial Occupation (%)	**-0.3706**	*0.2481*	-0.0055	-0.1781

C, Component. *Bolded if ≥|0.3015|. **Italicized if ≥|0.2015| and <|0.3015|.

The results of the second PCA using 2006 Census Profile data and 2005 taxfiler data are presented in Table 4. The first three principal components altogether account for 81.8% of the total variance, which is 3.3% more than the retained PCs in the first PCA using 2006 Census Profile data only. Compared to the first PCA using 2006 Census Profile data only, more of the variation has shifted to the first and second PCs (51.0% [+2.5%] and 20.0% [+3.6%], respectively) from the third PC (10.8% [-2.9%]). Additionally, PC_1 is influenced by most (9) of the 11 variables, while PC_3 is influenced mainly by a single variable, average monthly rent, and 2 lesser-weighted variables.

Table 4. Principal component analysis of 11 socioeconomic status indicator variables from 2006 Census Profile and 2005 taxfiler data for the census metropolitan area of Guelph.

Principal Component Eigenvalues					
Component	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**	**PC_W (%)**
PC_1	**5.613710**	**3.414840**	**0.5103**	**0.5103**	**62.39**
PC_2	**2.198860**	**1.014040**	**0.1999**	**0.7102**	**24.44**
PC_3	**1.184830**	**0.298320**	**0.1077**	**0.8179**	**013.17**
C_4	0.886505	0.414606	0.0806	0.8985	—
C_5	0.471899	0.258890	0.0429	0.9414	—
C_6	0.213008	0.022701	0.0194	0.9608	—
C_7	0.190307	0.074306	0.0173	0.9781	—
C_8	0.116001	0.037282	0.0105	0.9886	—
C_9	0.078719	0.053057	0.0072	0.9958	—
C_{10}	0.025662	0.005162	0.0023	0.9981	—
C_{11}	0.020500	—	0.0019	1	—

Indicator Variable Loadings (Eigenvector Correlations)				
Variable	**PC_1**	**PC_2**	**PC_3**	**Factor Score**
Median Family Income ($)	* **-0.3687**	0.1538	0.1220	-0.1764
Median Single Person Income ($)	**-0.3164**	*-0.2904*	*0.2619*	-0.2339
Low Income Families, After-tax (%)	**0.3697**	*0.2157*	-0.1286	0.2664
Low Income Unattached, After-tax (%)	** *0.2059*	**0.5395**	0.0046	0.2609
Lone Parent Families (%)	**0.4015**	0.0218	0.0437	0.2616
Single Mothers (%)	**0.3367**	-0.0199	*0.2628*	0.2398
Unemployment Rate, 15 years and over (%)	*0.2516*	*0.2733*	0.1483	0.2433
Low Education, 15 years and over (%)	0.1620	**-0.5230**	-0.1265	-0.0434
Average Home Value ($)	**-0.3290**	**0.3365**	-0.0002	-0.1231
Average Monthly Rent ($)	0.0054	0.0255	**0.8784**	0.1253
Managerial Occupation (%)	**-0.3354**	**0.3025**	-0.1395	-0.1537

C, Component. *Bolded if ≥|0.3015|. **Italicized if ≥|0.2015| and <|0.3015|.

The census data from 2006 produced a standardized SES index that increased consistently across most of the CTs. However, extreme SES Scores were seen at the lowest and highest levels of SES. Visual inspection revealed five distinct SES Score levels, which were further supported by mathematical cut-offs (Figure 1). The SES index produced from a combination of both 2006 Census and 2005 taxfiler data revealed clearer distinctions between SES levels, as well as a less extreme values at the high and low ends of the SES spectrum (Figure 2).

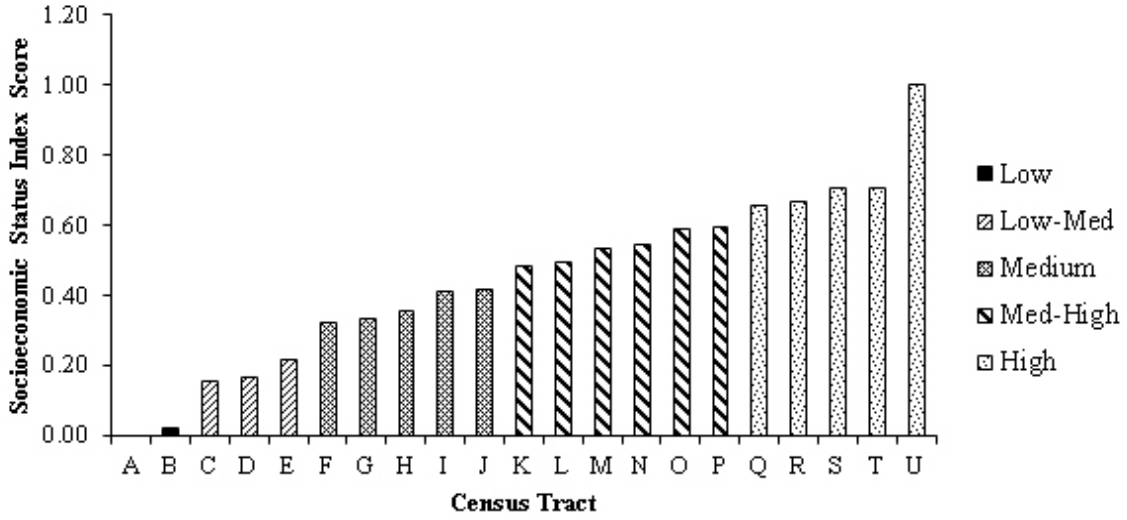

Figure 1. Socioeconomic status index distribution produced from the 2006 Census Profile dataset for the census metropolitan area of Guelph using all principal components with eigenvalues greater than 1.0 and that represent more than 10% of the variation.

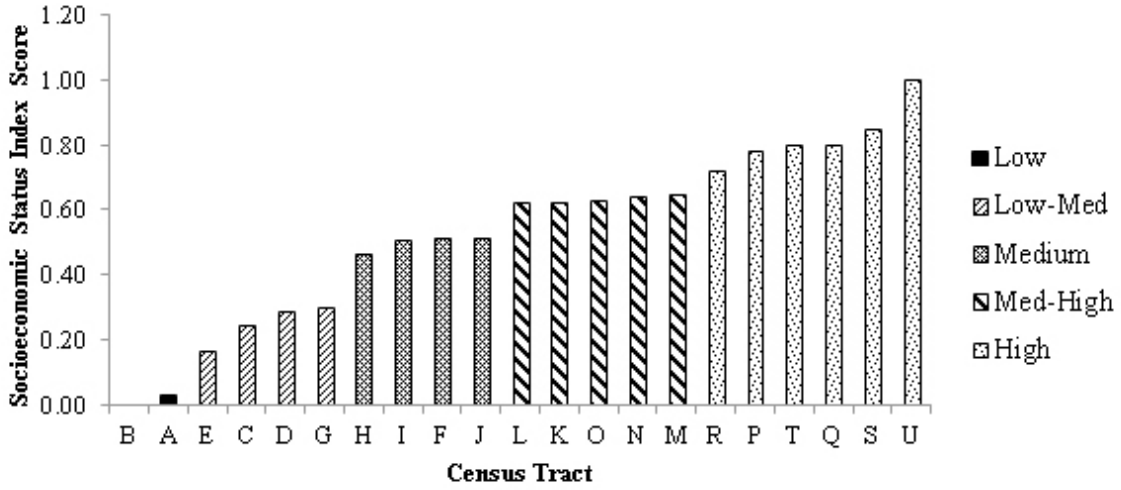

Figure 2. Socioeconomic status index distribution produced from the 2006 Census Profile and 2005 taxfiler datasets for the census metropolitan area of Guelph using all principal components with eigenvalues greater than 1.0 and that represent more than 10% of the variation.

SES classifications remained mostly consistent among CTs between the two data groups (Figure 3). Due to the overall consistency between both data groups and the shape of the SES index distribution, substituting taxfiler income variables for census income variables was deemed appropriate for creation of the 2011 SES index.

2006 Census Data Only			SES Level Change	2006 Census + 2005 taxfiler Data		
SES Level	Census Tract	SES Score		Census Tract	SES Score	SES Level
Low	A	0.0000		B	0.0000	Low
	B	0.0233		A	0.0295	
Low-Med	C	0.1557		E	0.1697	Low-Med
	D	0.1696		C	0.2476	
	E	0.2190		D	0.2887	
Medium	F	0.3241		**G**	**0.3015**	
	G	**0.3369**	**-1**	H	0.4663	Medium
	H	0.3581		I	0.5085	
	I	0.4105		F	0.5117	
	J	0.4200		J	0.5139	
Med-High	K	0.4843		L	0.6269	Med-High
	L	0.4941		K	0.6271	
	M	0.5367		O	0.6288	
	N	0.5435		N	0.6428	
	O	0.5904		M	0.6515	
	P	**0.5970**	**+1**	R	0.7226	High
High	Q	0.6562		**P**	**0.7850**	
	R	0.6667		T	0.7992	
	S	0.7050		Q	0.8038	
	T	0.7082		S	0.8522	
	U	1.0000		U	1.0000	

Figure 3. Changes to census tract socioeconomic levels for the census metropolitan area of Guelph after substituting 2006 Census income variables with 2005 taxfiler income variables using all principal components with eigenvalues greater than 1.0 and that represent more than 10% of the variation.

3) 2011 SES Index

The results of the third PCA using 2011 data from the Census, NHS, and taxfiler datasets are presented in Table 5. The total variance represented by the PCs is 73.03%, which is less than the two previous PCAs simply because only two components were retained in this case. Notably, more of the total variance shifted to the first PC compared to the two previous PCAs (48.49% to 51.03% to 53.03%). PC_1 is mainly influenced by median family income, proportion of low-income families, proportion of lone parent and single mother families, average home value and managerial occupation. Median single person income and proportion of low education have a slight influence on this PC. PC_2 is highly influenced by proportion of low-income individuals, single person income and unemployment rate, and somewhat influenced by proportion of low educated individuals over the age of 15, average home value, and average monthly rent. The datasets from 2011 produced a well-distributed standardized SES index with the clearest distinctions at the lower and higher ends of the scale (Figure 4).

Table 5. Principal component analysis of 11 socioeconomic status indicator variables from 2011 Census Profile, 2011 National Household Survey Profile and 2011 taxfiler datasets for the census metropolitan area of Guelph.

Principal Component Eigenvalues					
Component	**Eigenvalue**	**Difference**	**Proportion**	**Cumulative**	**PC_W (%)**
PC_1	**5.833060**	**3.632300**	**0.5303**	**0.5303**	**72.60**
PC_2	**2.200750**	**1.212010**	**0.2001**	**0.7303**	**27.40**
C_3	0.988738	0.194348	0.0899	0.8202	—
C_4	0.794390	0.338159	0.0722	0.8924	—
C_5	0.456231	0.185031	0.0415	0.9339	—
C_6	0.271200	0.088362	0.0247	0.9586	—
C_7	0.182837	0.046625	0.0166	0.9752	—
C_8	0.136213	0.024284	0.0124	0.9876	—
C_9	0.111929	0.089500	0.0102	0.9978	—
C_{10}	0.022428	0.020202	0.0020	0.9998	—
C_{11}	0.002227	—	0.0002	1	—

Indicator Variable Loadings (Eigenvector Correlations)			
Variable	**PC_1**	**PC_2**	**Factor Score**
Median Family Income ($)	* **-0.3885**	0.0826	-0.2595
Median Single Person Income ($)	** *-0.2841*	**-0.3261**	-0.2956
Low Income Families, After-tax (%)	**0.3315**	*0.2752*	0.3161
Low Income Unattached, After-tax (%)	0.1370	**0.6027**	0.2646

Lone Parent Families (%)	**0.3986**	-0.0284	0.2817
Single Mothers (%)	**0.3961**	-0.0295	0.2795
Unemployment Rate, 15 years and over (%)	0.0411	**0.4944**	0.1653
Low Education, 15 years and over (%)	*0.2763*	*-0.2496*	0.1322
Average Home Value ($)	**-0.3578**	*0.2293*	-0.1970
Average Monthly Rent ($)	-0.0643	*0.2190*	0.0133
Managerial Occupation (%)	**-0.3375**	0.1973	-0.1910

C, Component. *Bolded if ≥|0.3015|. **Italicized if ≥|0.2015| and <|0.3015|.

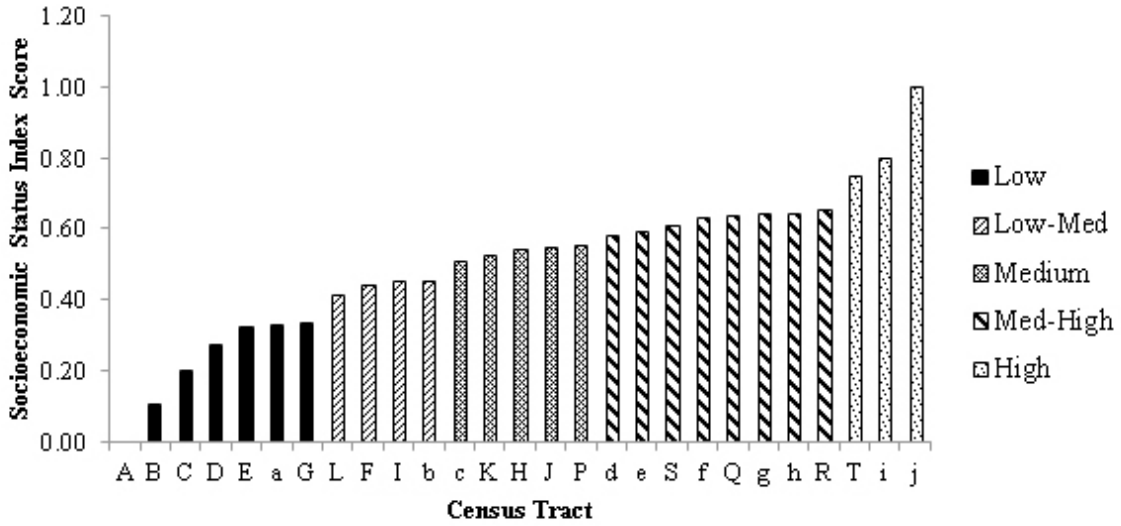

Figure 4. Socioeconomic status index distribution produced from the 2011Census Profile, 2011 National Household Survey Profile and 2011 taxfiler datasets for the census metropolitan area of Guelph using all principal components with eigenvalues greater than 1.0 and that represent more than 10% of the variation.

4) Internal Validation

The comparability among SES Scores across CTs between datasets was assessed by building SES indices for each dataset using only the first PC in the calculation. Using this method for the 2006 Census dataset produced an SES Score distribution with identical SES level distinctions as the original method, and CTs remained in the same SES levels.

This level of similarity was not found between the SES Score distributions when using the first PC from 2005 taxfiler and 2006 Census variables (Figure 5). Six CTs decreased in SES, while one CT increased one SES level. The transitory SES levels ('Low-Medium' and 'Medium-High') were comprised of fewer CTs while the lowest SES level ('Low') included

more CTs than when the SES Score was calculated using three components. Additionally, the distinctions between 'Medium', 'Medium-High', and 'High' SES were less pronounced.

	2006 Census + 2005 taxfiler Data (PCs >1.0)			2006 Census + 2005 taxfiler Data (PC1)		
SES Level	Census Tract	SES Score	SES Level Change	Census Tract	SES Score	SES Level
Low	B	0.0000		A	0.0000	Low
	A	0.0295		B	0.1575	
Low-Med	**E**	**0.1697**	**-1**	**C**	**0.2182**	
	C	**0.2476**	**-1**	**E**	**0.2362**	
	D	**0.2887**	**-1**	**D**	**0.2493**	
	G	0.3015		G	0.4024	Low-Med
Medium	H	0.4663		**F**	**0.4262**	
	I	0.5085		H	0.4891	Medium
	F	**0.5117**	**-1**	I	0.4946	
	J	0.5139		J	0.5010	
Med-High	**L**	**0.6269**	**-1**	**K**	**0.5355**	
	K	**0.6271**	**-1**	**L**	**0.5454**	
	O	0.6288		O	0.6014	Med-High
	N	**0.6428**	**+1**	M	0.6094	
	M	0.6515		**N**	**0.6815**	High
High	R	0.7226		P	0.6837	
	P	0.7850		S	0.6963	
	T	0.7992		Q	0.7190	
	Q	0.8038		R	0.7363	
	S	0.8522		T	0.8187	
	U	1.0000		U	1.0000	

Figure 5. Socioeconomic status scores for the census metropolitan area of Guelph produced from the substitution 2006 Census income variables with 2005 taxfiler income variables when using all principal components with eigenvalues greater than 1.0 and a proportion of explained variance greater than 10% ('PCs >1.0') versus using only the first principal component ('PC1').

Using the first PC resulted in even greater discrepancies in the 2011 SES Score distribution (Figure 6). Six CTs moved down one SES level while three CTs moved up one SES level. One CT moved down two levels from 'Medium-High' to 'Low-Medium'. This CT, 'S', was consistently in the 'Medium-High' or 'High' SES levels during previous calculations regardless of dataset or number of PCs used. The resulting distribution presented a larger distinction between 'Low' and 'Low-Medium' SES levels, with fewer clear distinctions throughout the higher levels of the scale. More of the CTs became classified as 'Low' or 'Low-Medium', with approximately half (14 or 51.8%) of the 27 CTs below the 'Medium' SES level.

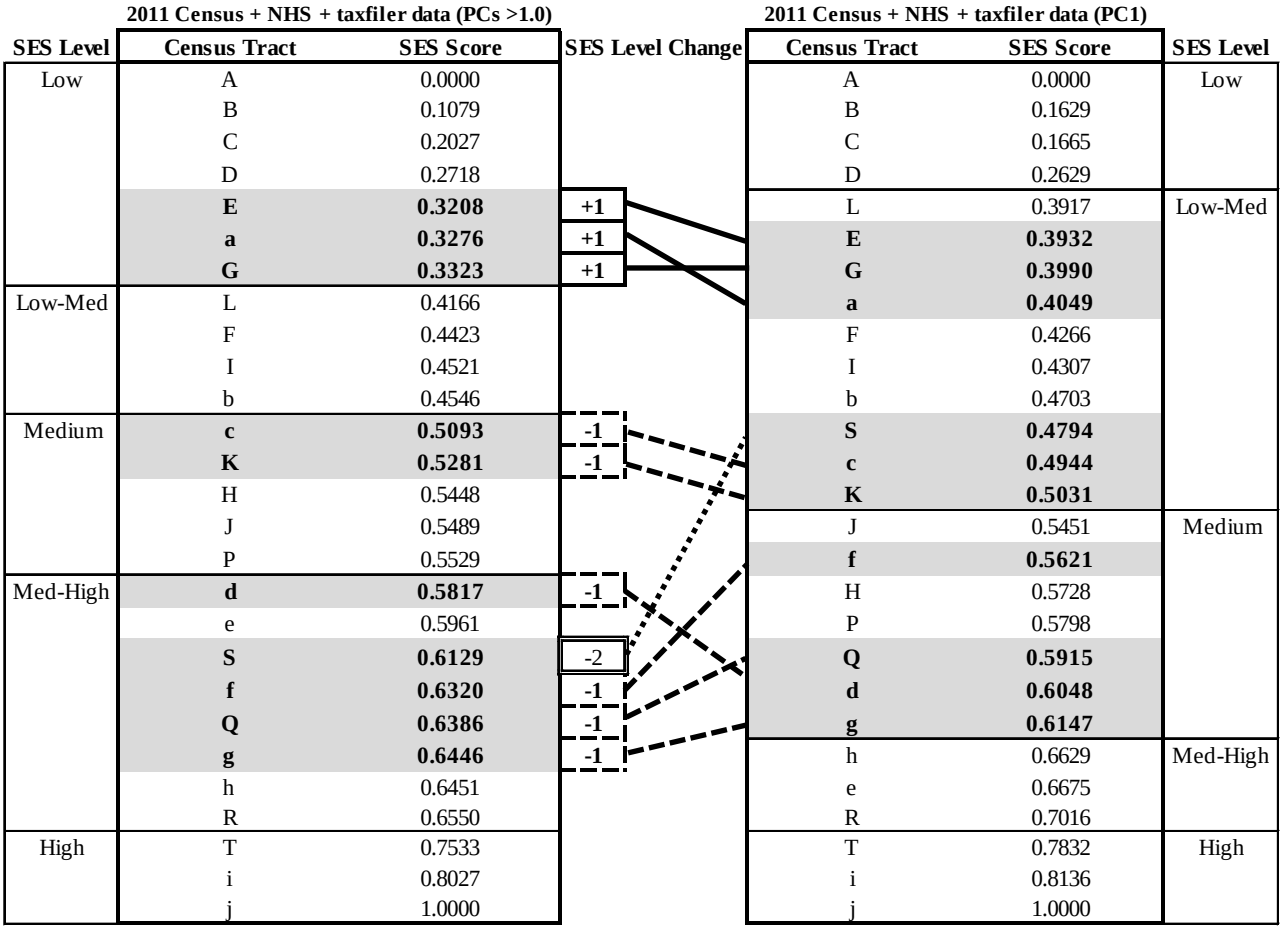

Figure 6. Socioeconomic status scores for the census metropolitan area of Guelph produced from a combination of the three 2011 datasets using all principal components with eigenvalues greater than 1.0 and a proportion of explained variance greater than 10% ('PCs >1.0') versus using only the first principal component ('PC1').

The second validation procedure excluded the seven non-income variables from the PCA (Table 6). Only the component met Kaiser's criterion of an eigenvalue >1.0 for all three datasets, and further consideration of proportion of explained variance suggested by Drackley *et al.* [4] was not pursued. The PC of the 2005 taxfiler data represented the greatest total variance (72.72%) while the PC of the 2006 Census data represented 63.25% of the total variance. The PC of the 2011 taxfiler data represented 69.16% of the total variance. The resulting variable loadings were approximately equally distributed amongst all variables for all datasets.

Table 6. A comparison of principal component analyses from three separate data groups (2006 Census Profile, 2005 taxfiler and 2011 taxfiler datasets) pertaining to the census metropolitan area of Guelph performed utilizing only the four income variables.

Source Dataset	Component	Eigenvalue	Difference	Proportion	Cumulative
2006 Census Profile	**PC_1**	**2.530160**	**1.741550**	**0.6325**	**0.6325**
	C_2	0.788604	0.206443	0.1972	0.8297
	C_3	0.582160	0.483082	0.1455	0.9752
	C_4	0.099079	—	0.0248	1.0000
2005 Taxfiler	**PC_1**	**2.908720**	**2.121430**	**0.7272**	**0.7272**
	C_2	0.787294	0.532456	0.1968	0.9240
	C_3	0.254838	0.205690	0.0637	0.9877
	C_4	0.049148	—	0.0123	1.0000
2011 Taxfiler	**PC_1**	**2.766560**	**1.929580**	**0.6916**	**0.6916**
	C_2	0.836980	0.485802	0.2092	0.9009
	C_3	0.351178	0.305902	0.0878	0.9887
	C_4	0.045277	—	0.0113	1.0000

Source Dataset	Indicator Variable	PC_1 Variable Loadings
2006 Census Profile	Median Family Income ($)	***0.5257**
	Median Single Person Income ($)	**0.538**
	Low Income Families, After-tax (%)	***-0.4864**
	Low Income Unattached, After-tax (%)	*-0.4445*
2005 Taxfiler	Median Family Income ($)	*-0.4477*
	Median Single Person Income ($)	**-0.5343**
	Low Income Families, After-tax (%)	**0.557**
	Low Income Unattached, After-tax (%)	*0.4516*
2011 Taxfiler	Median Family Income ($)	*-0.4593*
	Median Single Person Income ($)	**-0.5295**
	Low Income Families, After-tax (%)	**0.5505**
	Low Income Unattached, After-tax (%)	*0.4535*

C, Component. *Bolded if ≥|0.5|. **Italicized if ≥|0.4| and <|0.5|.

Note: 2006 Census Profile data produced signs opposite to the PCAs of the other two datasets. Signs were reversed for the 2006 PCA during SES Score calculations to ensure consistency.

The SES Indices produced from using only income variables maintained the CTs within one SES level across all three data groups as compared to the original calculations for the respective data groups (Figures 7, 8, & 9). However, by excluding non-income variables, fewer CTs were classified as 'Low' or 'Low-Medium' SES while more were classified as 'Medium-High' or 'High' SES. Additionally, the distinction between 'Low' SES and the other SES levels was much more pronounced, especially for 2005 taxfiler and 2011 taxfiler data groups.

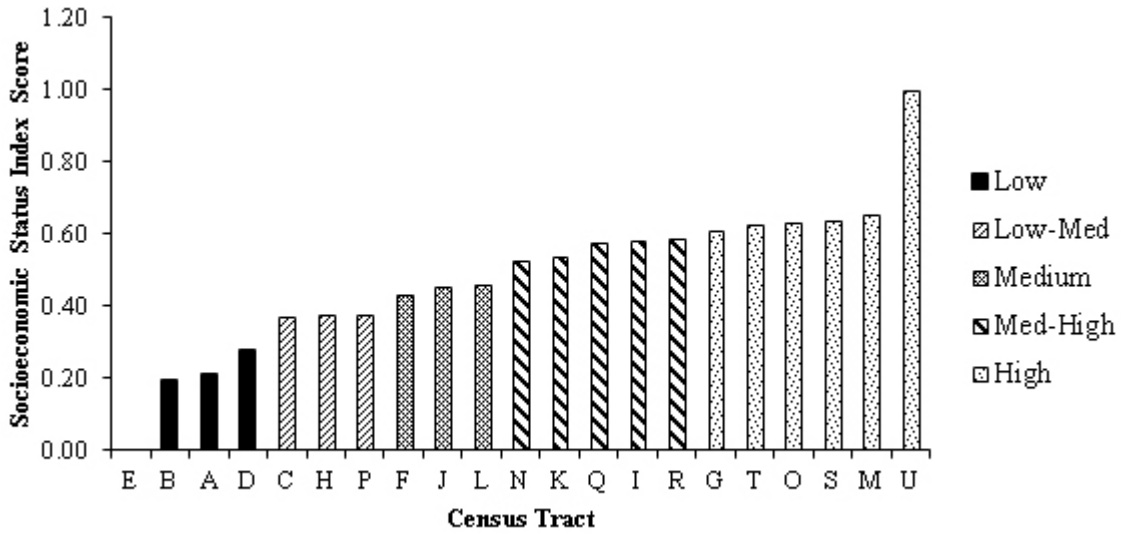

Figure 7. Socioeconomic status index distribution for the census metropolitan area of Guelph produced from the 2006 Census Profile dataset income variables and calculated using the first principal component only.

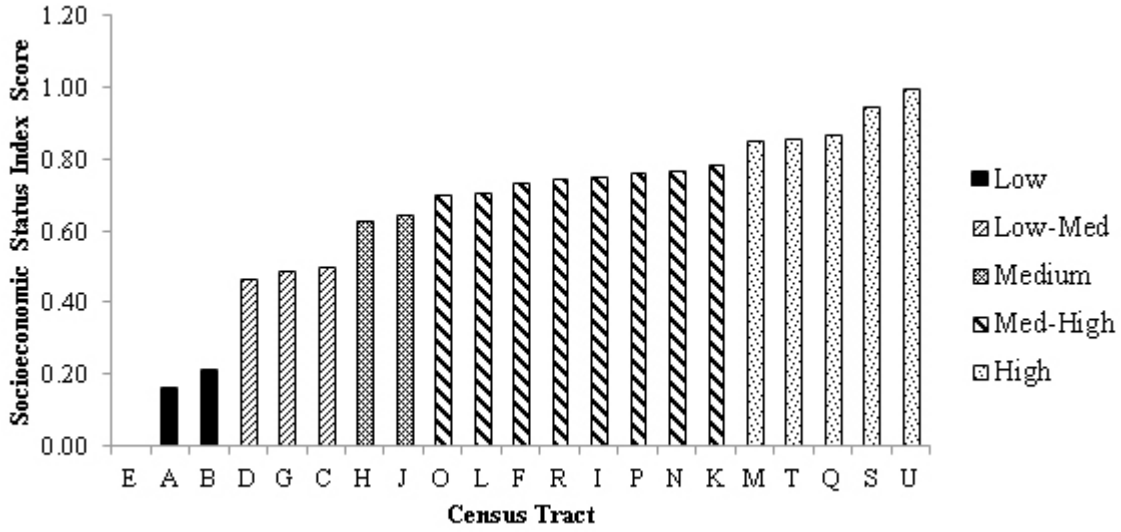

Figure 8. Socioeconomic status index distribution for the census metropolitan area of Guelph produced from the 2005 taxfiler income variables and calculated using the first principal component only.

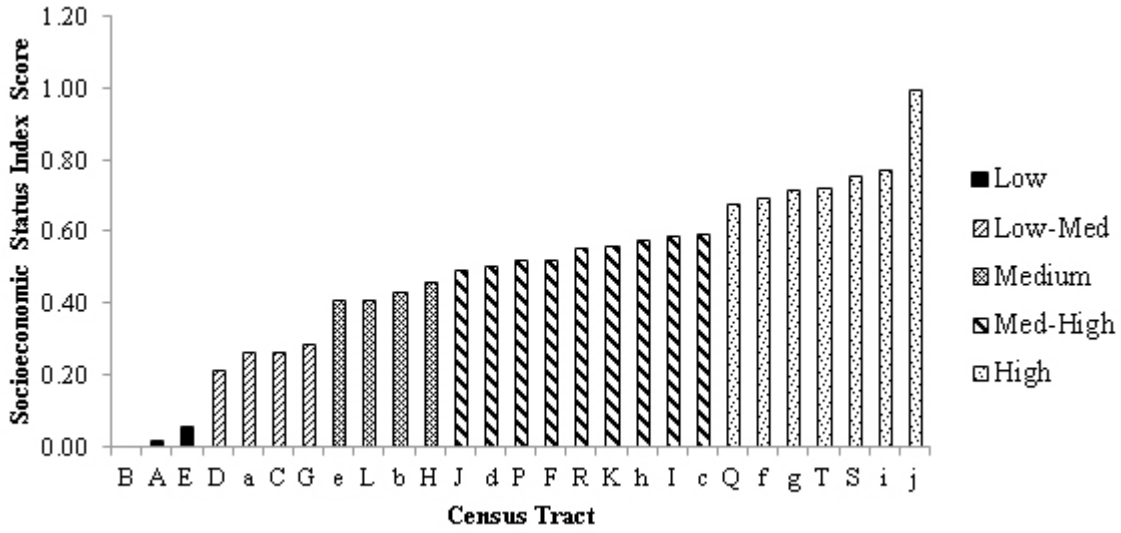

Figure 9. Socioeconomic status index distribution for the census metropolitan area of Guelph produced from the 2011 taxfiler income variables and calculated using the first principal component only.

DISCUSSION

Internal Validation

This report presented several SES indices across the CMA of Guelph produced using PCA on non-income variables indicative of SES and income variables derived from a novel data source. Internal validation showed that using the first PC for SES calculations resulted in skewed distributions with less pronounced distinctions between the SES levels. The SES index produced by the first PC method using 2011 data resulted in one CT dropping two SES levels, an indication that perhaps this method does not fully capture the dimensions of SES (Figure 6). Additionally, half of the CTs in 2011 using the first PC method were considered below the 'Medium' SES level, which would have severe implications on resource allocation for the support of public health in these areas. Several researchers have explained with confidence that the first PC is sufficient for calculating SES [5,7,9,10,17]. Interpreting additional components can be difficult and subjective, since the number of components produced is variable-dependent. Furthermore, some researchers deemed it unnecessary and potentially counterproductive to consider further components containing variable loadings that negate one another [9]. Vyas & Kumaranayake [9] considered a second component in their analysis but determined that it included a subset of variables not specific to a well-explained dimension and maintained that their first component representing wealth was sufficient for their SES index.

However, other researchers have explained that to adequately account for the complexity of SES, more PCs must be taken into consideration if they exceed an eigenvalue of 1.0 [6] or represent more than 10% of the variation [4]. These additional components represent more of the underlying variation and higher-order relationships between the variables used in the analysis. Krishnan [6] observed that up to five PCs were vital to represent economic, social, and cultural aspects of SES.

The present work further supports the inclusion of non-income variables in calculating SES. Using single variable measures like 'proportion of the population living in low income' presents a limited indication of an area's economic, social, cultural and health needs [4]. Social and geographical factors can significantly influence single variables, and therefore, a composite index better balances any changes incurred by single factors [6-8,11]. Additionally, measuring SES can be difficult in rural areas where measures of income do not consider long-term measures of wealth (e.g. self-subsistence agriculture) or assets that may better represent one's economic or social standing within their community [5,10]. When excluding non-income variables in the present analysis (Table 6), 'Medium-High' and 'High' SES levels were much more prevalent (Figures 7-9), which may be an over-representation of wealthy neighbourhoods. In contrast, the distribution of SES levels and clear distinctions between levels produced after including non-income variables further supports research of the past two decades into SES scales and indices [4-11].

Substituting Taxfiler Data

The present study has shown the comparability and superiority of tax filer to census income data. Including taxfiler data in the PCA resulted in a higher proportion of variance explained by the first few components, as well as an SES distribution with clear distinctions between levels and few extreme values. The use of readily available taxfiler data can assist in the confirmation of existing and identification of priority neighbourhoods by local public health units. A comparison to previous work by WDGPH revealed that due to differences in geographical level of analyses, some smaller areas of 'Low' SES in 2006 were suppressed within larger areas of 'High' SES in the present work of the same year. This was expected, since sources of income variables and PCA methods differed between the two works. The present method classified the majority of 'Low' areas from WDGPH in 2006 as either 'Low' or 'Low-Medium'.

In terms of identifying new priority areas in Guelph, the present analysis of 2011 data classified two areas as 'Low-Medium' SES that were previously 'High' in WDGPH's 2006 research. Additionally, a small area became 'High' SES in the present study where it was originally 'Low' in WDGPH's 2006 report. While this may be a result of geographical suppression, the methods presented here can act as part of a program evaluation tool for WDGPH as well as identify new communities that have since become disadvantaged and require new services.

Usefulness to Local Public Health Units

PCA can indicate which individual-level variables contribute the most to various dimensions of SES. This information can help local public health units prioritize existing programs and populations, as well as advise the development of new programs as necessary. The present analysis of the most recent available census and taxfiler data from 2011 produced two relevant PCs (Table 5). The first PC represents economic deprivation, as it is negatively influenced by family-related income variables, home value, and managerial occupation, and positively influenced by the proportion of low-income families, lone parent family structure, and the proportion individuals with low education. In other words, as family income increases and the proportion of the population with low education decreases, SES in that area would increase. This is in line with the current literature that identifies the first PC as economic conditions when using similar inputs [4-7,9,10,19]. The high weighting of the proportion of both lone parent families and single mothers indicates the importance of these factors in SES, and calls for further attention to these vulnerable groups by WDGPH programs like the *Triple P Initiative: Positive Parenting Program* [2].

The second PC is influenced mainly by single-person factors, with a slight contribution by the proportion of low-income families. Single person income contributes negatively to the second PC while the proportion of low-income individuals and unemployment rate contribute positively, suggesting that the second PC is a representation of unemployment conditions that have been associated with the SES and health of both individuals and populations [3].

This is similar to the finding by Drackley *et al.* [4], whose PCA attributed 'single renter' characteristics to the second PC. Current WDGPH initiatives to transition individuals out of poverty due to economic and employment situations, such as the *Getting Ahead* workshops and *Circles* support group, should continue to have a presence in the priority neighbourhoods identified using this method [2].

Three slightly influential variables in the second component presented directional effects on unemployment conditions that were counter-intuitive. The weightings of low education, average home value, and average monthly rent (-0.2496, 0.2293, and 0.2190, respectively) were approaching the minimum cut-off value of |0.2015|, which brings into question the significance of their influence on the second PC. According to the PCA performed using 2011 data (Table 5), an increased proportion of individuals with low education within the population are expected to reflect an improvement in the community's employment conditions. This association warrants further investigation, perhaps by evaluating career-training programs aimed at individuals who have not completed high school, as well as performing a similar PCA in different locales. The weighting of 'low education' was higher in the first principal component and weighted positively, suggesting that the relationship between low education and economic deprivation is stronger than with unemployment conditions.

Conversely, both average home value and average monthly rent positively contributed to individual-level unemployment conditions. Average home value weighted higher and positively in the first component, suggesting that it may be more influential on economic deprivation than unemployment conditions. In context of the second PC, average home value may be an indication of short-term or recent unemployment in which families or individuals are transitioning between SES levels.

Population demographics may play a role in the positive association between average monthly rent and unemployment conditions. Students who rent housing while attending the University of Guelph may not be employed during their studies, which would support the connection between monthly rent and unemployment rate. Since average monthly rent is weighted very low in both PCs, it may not be an appropriate indicator of SES in the city of Guelph, especially in the context of the community's demographics.

LIMITATIONS

There are some limitations to the present study. First, the addition of variables into the PCA can be subjective, which affects the quantity and weighting of components. This can be useful when exploring community-specific variables but may not be generalizable to larger areas. Second, as seen when comparing to previous work by WDGPH, using data at the census tract level may in fact suppress smaller areas of high priority within a classification of 'Medium' to 'High' SES that may deter further adjustments to existing public health programs in those areas. Finally, caution must be taken when inferring individual-level SES effects from the present aggregate-level data. Community turnover should to be assessed on an ongoing basis

using readily available individual-level variables to identify neighbourhoods that are consistently 'Low' and 'Low-Medium' SES.

CONCLUSION

Socioeconomic status and by extension, individual and population health, are influenced by many inter-related factors. The need for comprehensive approaches to health promotion and disease reduction that go beyond acute health care is becoming increasingly apparent. Identifying socioeconomic disparities between neighbourhoods is an important first step in assessing the level of disadvantage of communities, and the method presented here can be adapted to other locales for such a purpose.

The methods for developing an SES index presented in this paper support the use of PCA in assessing and ranking neighbourhoods using appropriate variables from that community. Further, the substitution of census income data with taxfiler data contributes to the current understanding of SES and population health. The present report supports a growing body of evidence that education, among other non-income variables, influences both familial and individual aspects of life, such information that should be used to support models such as the *Ontario Ministry of Children and Youth Services Strategic Framework* [28]. By improving SES measurement methods, a political shift may occur that moves the current system beyond poverty-reduction strategies into greater resource allocation to comprehensive programs targeting disadvantaged communities.

DECLARATIONS

The authors declare that they have no competing interests.

REFERENCES

1. Health Council of Canada (HCC). 2010. Stepping It Up: Moving the Focus from Health Care in Canada to a Healthier Canada. Toronto, ON. URL: http://www.healthcouncilcanada.ca/tree/2.40-HCCpromoDec2010.pdf

2. Wellington-Dufferin-Guelph Public Health (WDGPH). 2013. Addressing Social Determinants of Health in the City of Guelph: A public health perspective on local health, policy and program needs. Guelph, Ontario. URL: https://www.wdgpublichealth.ca/sites/default/files/wdgphfiles/sdoh-wdg-report-2013-for-web.pdf

3. Mikkonen J, Raphael D. 2010. Social Determinants of Health: The Canadian Facts. Toronto: York University School of Health Policy and Management. URL: http://www.thecanadianfacts.org/The_Canadian_Facts.pdf

4. Drackley A, Newbold KB, Taylor C. 2011. Defining Socially-Based Spatial Boundaries in the Region of Peel, Ontario, Canada. *Int J Health Geogr*. 10, 38-50. doi:http://dx.doi.org/10.1186/1476-072X-10-38. PubMed

5. Houweling TAJ, Kunst AE, Mackenbach JP. 2003. Measuring health inequality among children in developing countries: does the choice of the indicator of economic status matter? *Int J Equity Health*. 2, 8-20. doi:http://dx.doi.org/10.1186/1475-9276-2-8. PubMed

6. Krishnan V. 2010. Constructing an Area-based Socioeconomic Index: A Principal Components Analysis Approach. University of Alberta, Alberta, Canada. URL: http://www.cup.ualberta.ca/wp-content/uploads/2013/04/SEICUPWebsite_10April13.pdf

7. Messer LC, Laraia BA, Kaufman JS, Eyster J, Holzman C, et al. 2006. The Development of a Standardized Neighborhood Deprivation Index. *Journal of Urban Health: Bulletin of the New York Academy of Medicine*. 83(6), 1041-62. doi:http://dx.doi.org/10.1007/s11524-006-9094-x. PubMed

8. Oakes JM, Rossi PH. 2003. The measurement of SES in health research: current practice and steps toward a new approach. *Soc Sci Med*. 56, 769-84. doi:http://dx.doi.org/10.1016/S0277-9536(02)00073-4. PubMed

9. Vincent K, Sutherland J. 2013. A Review of Methods for Deriving an Index for Socioeconomic Status in British Columbia UBC Centre for Health Services and Policy Research. URL: http://healthcarefunding.ca/files/2013/04/Review-of-Methods-for-SES-Index-for-BC.pdf

10. Vyas S, Kumaranayake L. 2006. Creating Socio-economic Status Indices: How to use Principal Components Analysis. *Health Policy Plan*. 21(6), 459-68. http://heapol.oxfordjournals.org/cgi/content/abstract/21/6/459. PubMed http://dx.doi.org/10.1093/heapol/czl029

11. Yost K, Perkins C, Cohen R, Morris C, Wright W. 2001. Socioeconomic status and breast cancer incidence in California for different race/ethnic groups. *Cancer Causes Control*. 12, 703-11. doi:http://dx.doi.org/10.1023/A:1011240019516. PubMed

12. Fergusson DM, McLeod GFH, Horwood LJ. 2015. Leaving school without qualifications and mental health problems to age 30. *Soc Psychiatry Psychiatr Epidemiol*. 50, 469-78. doi:http://dx.doi.org/10.1007/s00127-014-0971-4. PubMed

13. Hankivsky O. 2008. Cost Estimates of Dropping Out of High School in Canada. Canadian Council on Learning. URL: http://www.ccl-cca.ca/pdfs/OtherReports/CostofdroppingoutHankivskyFinalReport.pdf

14. Vaughn MG, Beaver KM, Wexler J, DeLisi M, Roberts GJ. 2011. The Effect of School Dropout on Verbal Ability in Adulthood: A Propensity Score Matching Approach. *J Youth Adolesc*. 40(2), 197-206. doi:http://dx.doi.org/10.1007/s10964-009-9501-1. PubMed

15. Zajacova A, Everett BG. 2014. The Nonequivalent Health of High School Equivalents. *Soc Sci Q*. 95(1), 221-38. doi:http://dx.doi.org/10.1111/ssqu.12039. PubMed

16. Public Health Agency of Canada (PHAC). 2003. What makes Canadians healthy or unhealthy. URL: http://www.phac-aspc.gc.ca/ph-sp/determinants/index-eng.php

17. Filmer D, Pritchett LH. 2001. Estimating wealth effects without expenditure data – Or tears: An application to educational enrollments in States of India. *Demography*. 38(1), 115-32. doi:10.1353/dem.2001.0003. PubMed

18. Primpas I, Tsirtsis G, Karydis M, Kokkoris GD. 2010. Principal component analysis: Development of a multivariate index for assessing eutrophication according to the European water framework directive. *Ecol Indic*. 10, 178-83. doi:http://dx.doi.org/10.1016/j.ecolind.2009.04.007.

19. Messer LC, Jagai JS, Rappazzo KM, Lobdell DT. 2014. Construction of an environmental quality index for public health research. *Environ Health*. 13, 39-61. doi:http://dx.doi.org/10.1186/1476-069X-13-39. PubMed

20. Statistics Canada. 2010. 2011 Census of Population, P.C. 2010-1077. *Can Gaz, I*. 144(34), 2257-307.

21. Statistics Canada. 2014. 2011 Census questionnaire. Statistics Canada. URL: http://www12.statcan.gc.ca/census-recensement/2011/ref/gazette-eng.cfm

22. Canadian Council on Social Development (CCSD). 2015. Community data program. Canadian Council on Social Development. URL: http://ccsd.ca/index.php/enable/community-data-program

23. Statistics Canada. 2012. Census tract (CT). Census Dictionary, catalogue no. 98-301-X. URL: https://www12.statcan.gc.ca/census-recensement/2011/ref/dict/geo013-eng.cfm

24. Statistics Canada, 2013. Hierarchy of standard geographic units. Illustrated Glossary, catalogue no. 92-195-X.

25. Statistics Canada. 2009. Low income cut-offs for 2008 and low income measures for 2007. Income Research Paper Series, catalogue no. 75F0002M, no. 002. URL: http://www.statcan.gc.ca/pub/75f0002m/75f0002m2009002-eng.pdf

26. Hair JF. Jr., Anderson, R.E., Tatham, R.L., Black, W.C., 1998. *Multivariate Data Analysis*, (5th Edition). Prentice Hall, Upper Saddle River, NJ.

27. Raubenheimer J. 2004. An item selection procedure to maximise scale reliability and validity. *SA J Ind Psychol*. 30(4), 59-64. doi:http://dx.doi.org/10.4102/sajip.v30i4.168.

28. Ontario Ministry of Children and Youth Services (OMCYS). 2008. Realizing Potential: Our Children, Our Youth, Our Future. URL: http://www.children.gov.on.ca/htdocs/English/documents/about/StrategicFramework.pdf

Appendix A

Datasets retrieved from the Canadian Council on Social Development's Community Data Program [22].

Census Datasets	Catalogue Number
Census Profile, 2006 *Data Provider:* Statistics Canada *Survey Number:* 3901 *Release Date:* May 1, 2008	94-581-x2006001, 94-581-x2006002, 94-581-x2006003, 94-581-x2006004, 94-581-x2006005, 94-581-x2006008
Census Profile, 2011 *Data Provider:* Statistics Canada *Survey Number:* 3901 *Release Date:* October 24, 2012	98-314-x2011006, 98-314-x2011007, 98-314-x2011008, 98-314x-2011009, 98-314-x2011010, 98-314-x2011011, 98-314-x2011012, 98-314-x2011013, 98-314-x2011014, 98-314-x2011015, 98-314-x2011052
NHS Profile, 2011 *Data Provider:* Statistics Canada *Survey Number:* 5178 *Release Date*: September 11, 2013	99-004-x2011015, 99-004-x2011016, 99-004-x2011017, 99-004-x2011018, 99-004-x2011019
Taxfiler (T1FF) Datasets	**Contents**
F01: Total Income Summary Table *Data Provider:* Statistics Canada *Years Obtained*: 2005, 2011	Table F-1 Family data - Summary, 2005
	Table F-1 Family data - Summary, 2011
F04: Total Income by Family Type *Data Provider:* Statistics Canada *Years Obtained*: 2005, 2011	Table F-4A Family data - Distribution of Total Income of Couple Families by Age of Older Partner, 2005 Table F-4B Family data - Distribution of Total Income of Lone-Parent Families by Age of Parent, 2005 Table F-4C Family data - Distribution of Total Income of Person not in Census Families by Age, 2005
	Table F-4A Family data - Distribution of Total Income of Couple Families by Age of Older Partner, 2011 Table F-4B Family data - Distribution of Total Income of Lone-Parent Families by Age of Parent, 2011 Table F-4C Family data - Distribution of Total Income of Person not in Census Families by Age, 2011
F17: Before Tax Low-Income *Data Provider:* Statistics Canada *Years Obtained*: 2005, 2011	Table F-17 Family data - Low income (based on before-tax low income measures, LIMs), 2005
	Table F-17 Family data - Low income (based on before-tax low income measures, LIMs), 2011
F18: After Tax Low-Income *Data Provider:* Statistics Canada *Years Obtained*: 2005, 2011	Table F-18 Family data - After-tax low income (based on after-tax low income measures, LIMs), 2011
	Table F-18 Family data - After-tax low income (based on after-tax low income measures, LIMs), 2005