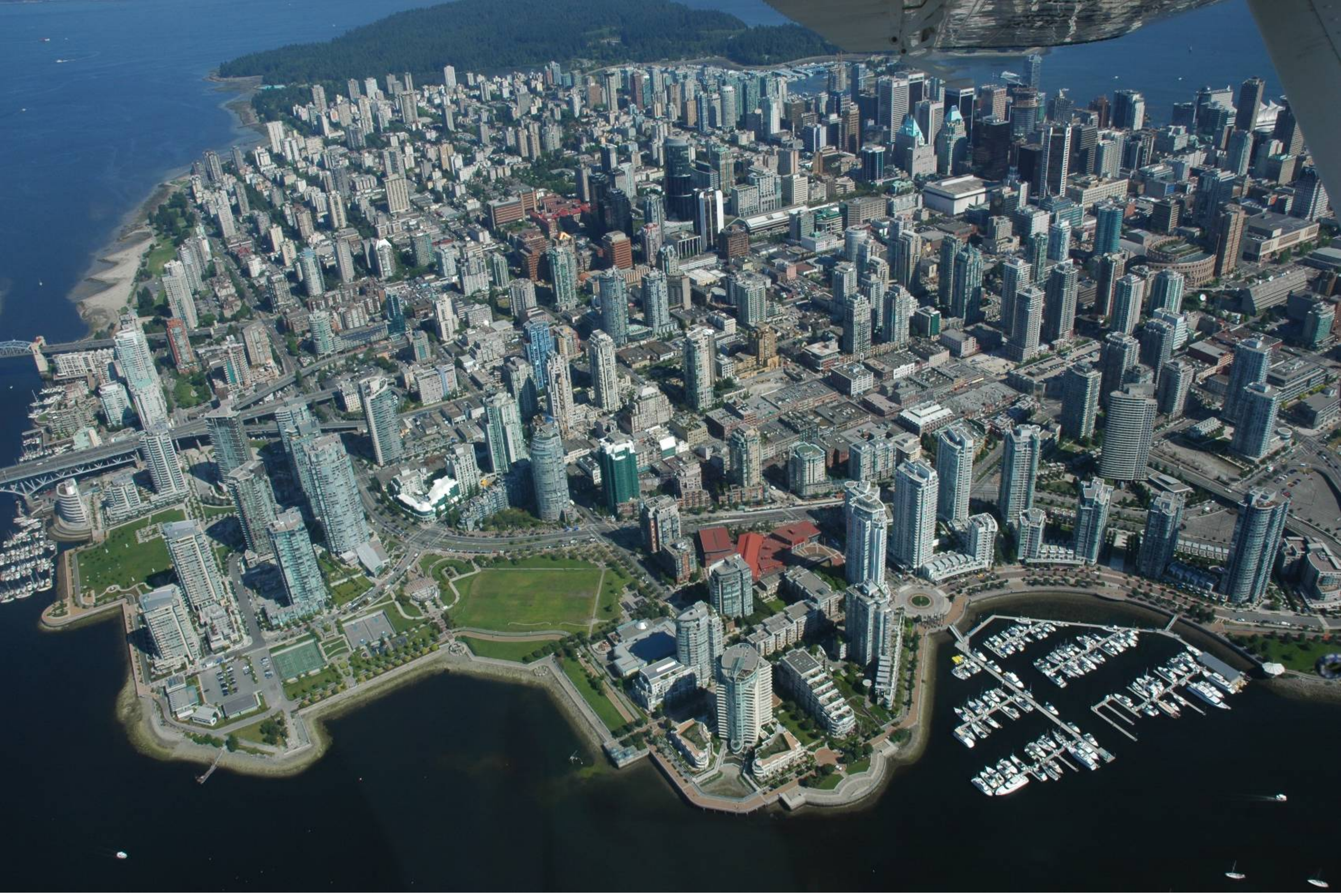

Figure 1. Central Vancouver, June 2006 (Elvin Wyly).

Mapping Vancouver's Evolving Social Mosaic

Project background paper
Geography 350, *Introduction to Urban Geography*
November 7, 2012
Elvin Wyly[1]

> "The societal and spatial changes of the past 20 years seem sufficiently important to have produced new patterns of social differentiation in western cities. In general terms, the changes are summarized in the term post-industrial society, ... and can be specifically linked to the increasing complexity of family life, the changing position of women ... and minority groups. ... It is time to explore the effect of these changes upon the contemporary urban social differentiation of western cities in a multivariate, not a single variable context ...."
>
> Wayne Davies and Robert Murdie[2]

[1] I am grateful to Markus Moos and Cory Dobson for help in assembling parts of the databases used in this background paper.

[2] Wayne K.D. Davies and Robert A. Murdie (1991). "Consistency and Differential Impact in Urban Social Dimensionality: Intra-Urban Variations in the 24 Metropolitan Areas of Canada." *Urban Geography* 12(1), 55-79, quote from p. 46.

Social area analysis was an approach developed in the 1950s to analyze the varied social and demographic characteristics of neighborhoods inside cities. The approach emphasized three fundamental dimensions of urban social space: economic status, family status, and ethnic status.

Much of the history of urban geography is the history of attempts to map, describe, and explain the extraordinary complexity of the social spaces of the city. More than half a century ago, analysts began to devise systematic ways of summarizing this social complexity down to a smaller number of general dimensions that could help distinguish different kinds of communities in the metropolis. Early on, much of this "social area analysis" – launched by Eshref Shevky and Wendell Bell in a book with the same title[3] -- was inspired by the powerful idea that although there may be an unlimited number of ways of describing and measuring different neighborhoods in the city, the fundamental essence of urbanization could be distilled down into three dimensions: ***economic status***, which expressed the contrasts between wealthy, middle-class, and poorer parts of the city; ***family status***, which distinguished areas with large families and children, as opposed to districts dominated by singles and/or elderly persons; and ***ethnic status***, which captured the way that different racial and ethnic groups tend to be concentrated in particular communities. After a wave of research in which these ideas motivated the study of common patterns across many different cities, however, many researchers began to believe that cities were growing more complex, and that the increasing complexity of any particular city was in large part a reflection of the changes underway in its society. At the same time, methodological and technological advances made it possible to measure various aspects of cities in ever more fine-grained detail. This led to a new wave of research that sought to document the extraordinary variety of neighborhood social patterns, and the rapid pace of spatial changes brought on by dramatic economic, cultural, and political transformation. This kind of research was dubbed "urban factorial ecology," because it combined a technique known as factor analysis with a theoretical tradition -- human ecology -- that was part of the influential Chicago School of Sociology. For a time, this research came to dominate the field of urban geography as well as urban sociology and parts of urban planning. Eventually, though, many analysts turned to other questions and other methods. Many began to *assume* that the urban mosaic was always in flux, always changing dramatically in line with sweeping societal upheaval. Yet others would argue that complexity and specificity should not distract us from general, common divisions that still endure.

Urban factorial ecology combined a statistical technique -- factor analysis -- with the theories of human ecology that were part of the dominant Chicago School of Sociology.

How can we measure and describe the complex social mosaic of the city? Is the city best understood in terms of a coming together, a general unity that brings together many different

[3] Eshref Shevky and Wendell Bell (1955). *Social Area Analysis.* Stanford: Stanford University Press.

people into coherent local communities that can be distinguished on the basis of economic, family, and racial/ethnic relations? Or is the city more fragmented, complex, and hard to understand as any kind of integrated community? These are among the most difficult, provocative, and valuable questions in the study of cities. There are no indisputably correct or incorrect answers, and there are quite compelling, persuasive arguments for various explanations and interpretations. In this project, my goal is not to tell you *what* to think about the changing social fabric of the city. Instead, I want to provide you with a few tools that will show you one way of *how* to think about the social space of the city.

Specifically, I describe a set of methods that will help you to analyze many different aspects of several hundred neighborhoods in the Vancouver metropolitan area. These methods are used not just in urban geography, but throughout many parts of the social sciences, and in private industry as well. The specifics of the methods – principal components analysis, and factor analysis – can get quite detailed and specialized. But don't let the details frighten you. First, the general purpose of all these detailed techniques is really quite simple: how do we take a large number of different ways of measuring things and understand how they relate to one another, and to see if they are in fact measuring the same tendencies? Second, this project does **not** require you to actually do a principal components analysis or factor analysis. I've already done them for you. You simply need to read through this background paper to learn enough about the approach so that you know how to interpret the results. Your job in this project is simple: choose any combination from among those many different neighborhoods in the Vancouver region, and then tell a story about urban social change that draws in some way on the traditions of social area analysis and factorial ecology. You can choose a general path (studying general tendencies across many different neighborhoods) or a more specific path (examining a few neighborhoods in detail and explaining how they fit into the broader social mosaic).[4]

"Placing" PCA and Factor Analysis: Geography and Method

Some of the best studies of the complexity of urban social space rely on a set of methods known as principal components analysis (PCA) and factor analysis. Although the methodological purists might take exception to the metaphor, it is easiest to grasp the differences between these two approaches by imagining factor analysis as principal components on steroids. This section provides a review of where the approaches came from and how they came to be used in urban research; then in the next section we consider a simple illustration of how the technique works.

[4] In addition to this background paper, I recommend the following sources for additional information on urban factorial ecology: Robert A. Murdie and Carlos Teixeira (2006). "Urban Social Space." Chapter 9 in Trudi Bunting and Pierre Filion, eds., *Canadian Cities in Transition: Local Through Global Perspectives*. Don Mills, ON: Oxford University Press, 154-170; Elvin K.Wyly (1999). "Continuity and Change in the Restless Urban Landscape." *Economic Geography* 75(4), 309-338; and Paul Knox and Linda McCarthy (2005). *Urbanization, Second Edition.* Upper Saddle River, NJ: Pearson/Prentice-Hall, pp. 311-339.

Principal components analysis was devised in the early years of the twentieth century.[5] Factor analysis was developed around the same time, using similar mathematical procedures, in the field of educational psychology.[6] Educational researchers often encountered the problem of many different variables (say, scores on various tests, or grades in specific subjects) that were all attempting to measure different aspects of the same underlying construct (aptitude, achievement, or, much more controversially, a fundamental, underlying 'intelligence'). Principal components analysis and factor analysis provided systematic ways of determining how separate, multiple measures – those scores and grades on different tests and subjects – related to one another, and hinted at the contours of an underlying, latent dimension or factor. "Underlying, latent dimension:" keep an eye out for terms like this in any scholarship that makes use of principal components analysis or factor analysis. The terms hint at the kind of thinking involved: in many areas of research we have lots and lots of simple measures or indicators. But often they don't seem to capture the full complexity of the *concept* we are trying to describe, analyze, and explore.

Principal components analysis and factor analysis provide a way of measuring how multiple indicators relate to one another -- and how they reflect an underlying latent dimension or factor. These approaches can help us to understand how single indicators (income, employment, education, etc.) reflect an underlying, multi-faceted concept, like urban class inequality.

Principal components analysis and factor analysis took off across many of the social sciences in the 1950s and 1960s. The method became especially popular in urban geography and urban sociology, in the era when analytical urban geography was emerging as a forceful movement advocating the use of quantification to uncover order amidst complex spatial patterns. When confronted with an especially complex spatial pattern, the geographer could use principal components analysis or factor analysis to sift through the complexity to uncover the latent structure of relations in a place.[7] Through the 1960s and 1970s, "factorial ecology" became all the rage in studies of neighborhood social, economic, and housing conditions in big cities around the world – *or at least around those parts of the world where it was possible to get detailed information about conditions in city neighborhoods.*[8] But soon the movement generated a

[5] K. Pearson (1901). "On Lines and Planes of Closest Fit to Systems of Points in Space." *Philosophical Magazine* 6(2), 559-572. See also H. Hotelling (1933). "Analysis of a Complex of Statistical Variables into Principal Components." *Journal of Educational Psychology* 24, 417-441, 498-520.

[6] C. Spearman (1904). "General Intelligence Objectively Determined and Measured." *American Journal of Psychology* 15, 201-293.

[7] See Peter Gould (1967). "On the Geographical Interpretation of Eigenvalues." *Transactions of the Institute of British Geographers*, Winter, 53-86. See also D. Michael Ray (1969). "The Spatial Structure of Economic and Cultural Differences: A Factorial Ecology of Canada." *Papers in Regional Science* 23(1), 7-23.

[8] For contributions and assessments in the urban factorial ecology literature, I recommend these sources: Berry, Brian J.L., and Horton, Frank E. (1970). *Geographic Perspectives on Urban Systems*. Englewood Cliffs, NJ: Prentice-Hall. Berry, Brian J.L., and Kasarda, John D. (1977). *Contemporary Urban Ecology*. New York: Macmillan. Davies, Wayne K.D. (1984). *Factorial Ecology*. Aldershot: Gower Press. Johnston, R.J. (1984). *City and Society: An Outline for Urban Geography*. London: Hutchinson. On the more general point of shifting

After a period of dominance, there was a backlash against the methods of urban factorial ecology. Most urban geographers turned to other methods and theories. But factorial ecology became enormously popular in private industry, and it is widely used in retail site selection, consumer segmentation, and strategic marketing.

backlash, and in recent years most geographers studying globalization, economic restructuring, and widespread social and cultural transformations have chosen to use the analytical tools of humanism, phenomenology, structuralism, and poststructuralism.

But even as fewer urban geographers used the techniques of factorial ecology, the approach became ever more pervasive in some quarters of the social sciences, while also proliferating in an exploding and lucrative industry movement. The "geodemographic marketing industry" consists of an ensemble of technologies, companies, and practices devoted to creating and analyzing locationally-referenced information about individual consumer behavior in order to inform decisions about retail locations, marketing tactics, inventory management, and a variety of other aspects of consumption landscapes; principal components analysis and factor analysis are key tools in this industry.[9] Factor analysis is also now widely used in facial recognition software and other automated algorithms used for the surveillance of public places.[10]

For all of these reasons, it is extremely helpful to understand some of the basic steps involved in principal components analysis and factor analysis.

A Simple Geometric View of Principal Components Analysis

Consider a simple dataset with two *variables* measured across fifty *observations* (see Table 1). The observations are census tracts – zones within urban areas that are defined, for the purposes of the Canadian Census of population and housing, to provide information on the changing social

questions and techniques in the urban literature, consult Livingstone's chapter in *The Geographical Tradition*, or, for the one of the most comprehensive urban bibliographies published in the last decade, see Harris, Chauncy D. 1995. "'The nature of cities' and urban geography in the last half century." *Urban Geography* 18(1): 15-35.

[9] For a recent sample of some of the fusions of factor analysis, cluster analysis, and consumer-behavioral data, see Peter Duchessi, Charles M. Schaninger, and Thomas Nowak (2004). "Creating Cluster-Specific Purchase Profiles from Point-of-Sale Scanner Data and Geodemographic Clusters: Improving Category Management at a Major U.S. Grocery Chain." *Journal of Consumer Behavior* 4(2), 97-117. For critical evaluations of this wave of geodemographic innovation, see Jon Goss (1995). "We Know Who You Are and We Know Where You Live: The Instrumental Rationality of Geodemographic Systems." *Economic Geography* 71(2), 171-198. For updates and extensions, see Martin Dodge and Rob Kitchin (2005). "Codes of Life: Identification Codes and the Machine-Readable World." *Environment and Planning D: Society and Space* 23, 851-881.

[10] For a frightening sample of some of the kinds of research in this area, see Siegfried Ludwig Sporer, Barbara Trinkl, and Elena Guberova (2007). "Matching Faces: Differences in Processing Speed of Out-Group Faces by Different Ethnic Groups." *Journal of Cross-Cultural Psychology* 38(4), 398-412. For a critical evaluation of these practices authored by an alum of UBC's Urban Studies and Journalism programs, see Mitchell Gray (2003). "Urban Surveillance and Panopticism: Will We Recognize the Facial Recognition Society?" *Surveillance and Society* 1(3), 314-330.

and housing mosaic of various neighborhoods. To keep this illustration simple, I've just selected fifty census tracts for a part of central Vancouver, from Kitsilano to Mount Pleasant and including the entire downtown peninsula (see Figure 2). For each census tract, we have two variables from the 2001 Census: the median annual income of households, expressed as a ratio to the overall figure for the metropolitan area,[11] and the proportion of private dwellings occupied by residents who own their homes. Census Tract 68, for instance, has a median household income that is 76.01 percent of the value for the entire metropolitan area, and a homeownership rate of 15.68 percent; this tract includes the older apartment buildings West of Denman Street next to Stanley Park.

For many purposes, it is helpful to measure things in terms of whether they are above average or below average. We can do this if we subtract each value from the overall mean for that variable, giving us what is known as "mean-corrected" data. For our small dataset, the mean-corrected data indicate that Tract 68 in the West End is a little bit below the mean for our set of fifty tracts in Central Vancouver, both in terms of income (-0.0629), and homeownership (-.1879). By contrast, Tract 59.03, the North Shore of False Creek, stood well above the mean in terms of income (.4006) and ownership (.1321). The many differences between these neighborhoods can be summarized in the ***variance*** for our small dataset. The variance is just what it sounds like: a measure of how much a set of values vary from the mean. The variance is calculated as *the sum of the squared deviation of each value from the mean, divided by the number of observations*.[12] The square root of the variance is equal to the **standard deviation**, which can be understood as the *average distance of each observation from the mean*. For our small glimpse of central Vancouver, our two-variable dataset has a total variance of 0.1187 (0.0849 for the first variable, 0.038 for the second variable).

[11] The median household income for the entire Vancouver Census Metropolitan Area in 2001 was $49,940.

[12] Using the sum of the squared deviations is a convenient way of getting around an annoying problem: if we simply add up all the differences between each observation and the mean, it will always by definition sum to zero. Squaring the differences solves this problem easily.

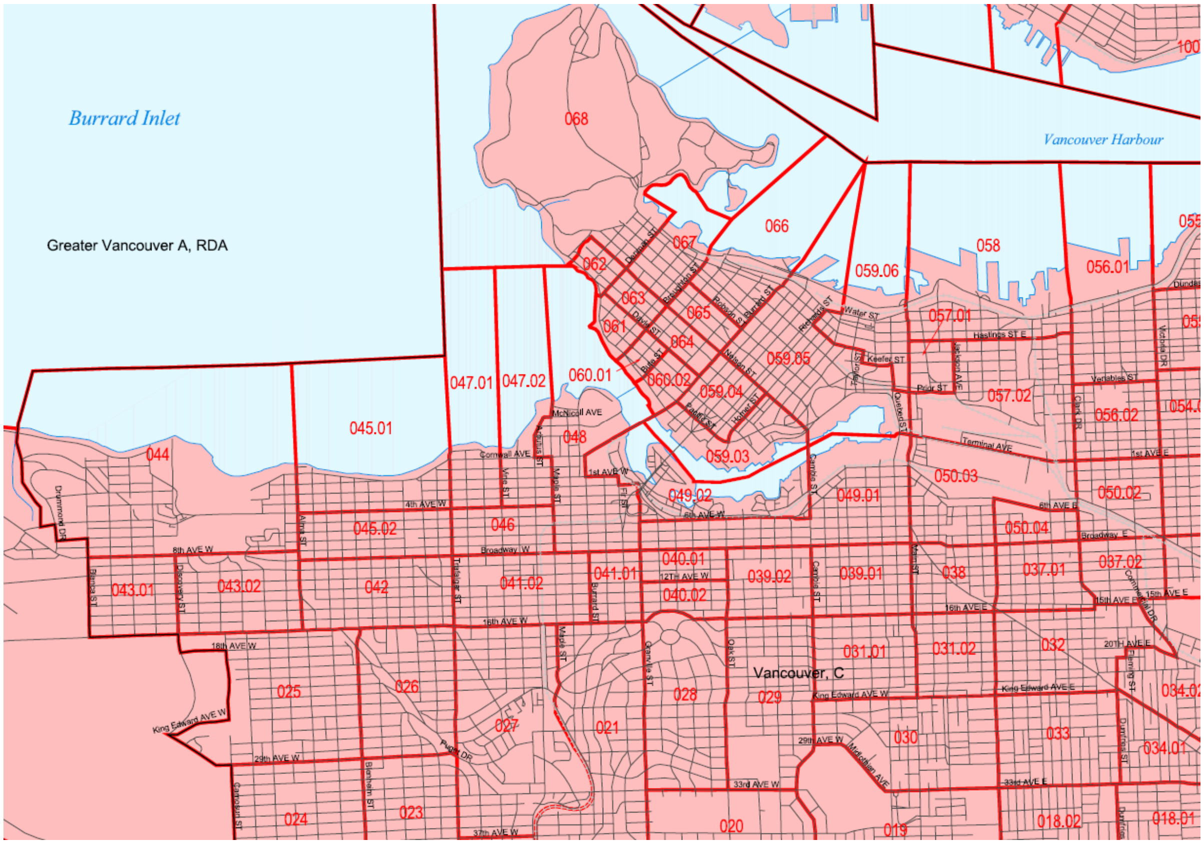

Figure 2. Census Tract Boundaries in Central Vancouver, 2001. Data Source: Statistics Canada (2002). *Census Tract Reference Maps, Reference Guide. Catalogue No. 92F0145GIE.* Ottawa: Statistics Canada.

Table 1. Income and Homeownership for Census Tracts in Central Vancouver, 2001. Data Source: Statistics Canada (2003). *Electronic Profiles, Census Metropolitan Areas, Tracted Census Agglomerations, and Census Tracts, 2001 Census*. Release 95F0495XCB2001005. Ottawa: Statistics Canada.

Census Tract	Median household income ratio (X1)	Mean-corrected	Homeownership Rate (X2)	Mean-corrected
38	0.655	-0.168	0.246	-0.099
39.01	0.764	-0.059	0.213	-0.132
39.02	0.869	0.046	0.363	0.018
40.01	0.815	-0.008	0.137	-0.208
40.02	0.852	0.029	0.208	-0.136
41.01	0.886	0.063	0.278	-0.067
41.02	1.179	0.356	0.558	0.213
42	1.097	0.273	0.541	0.197
43.01	1.394	0.571	0.553	0.208
43.02	1.708	0.885	0.728	0.384
44	1.227	0.404	0.581	0.236
45.01	1.262	0.439	0.451	0.106
45.02	1.077	0.254	0.436	0.092
46	0.844	0.021	0.291	-0.053
47.01	0.941	0.118	0.325	-0.019
47.02	0.948	0.125	0.244	-0.101
48	0.961	0.138	0.359	0.014
49.01	1.145	0.322	0.502	0.157
49.02	1.054	0.231	0.453	0.108
50.02	0.553	-0.270	0.208	-0.137
50.03	0.606	-0.217	0.343	-0.002
50.04	0.516	-0.307	0.285	-0.060
51	0.848	0.025	0.623	0.279
52.01	0.696	-0.127	0.543	0.199
52.02	0.858	0.035	0.596	0.252
53.01	0.920	0.097	0.606	0.262
53.02	1.079	0.256	0.647	0.302
54.01	0.916	0.093	0.529	0.185
54.02	0.944	0.121	0.718	0.374
55.01	0.600	-0.223	0.342	-0.003
55.02	0.657	-0.166	0.350	0.005
56.01	0.435	-0.388	0.137	-0.208
56.02	0.682	-0.141	0.238	-0.107
57.01	0.275	-0.548	0.045	-0.299
57.02	0.369	-0.454	0.223	-0.122
58	0.217	-0.606	0.047	-0.298
59.03	1.224	0.401	0.477	0.132
59.04	0.608	-0.215	0.288	-0.056
59.05	0.816	-0.007	0.431	0.086
59.06	0.201	-0.622	0.077	-0.268
60.01	0.743	-0.080	0.143	-0.202
60.02	0.707	-0.117	0.181	-0.164
61	0.707	-0.116	0.099	-0.246
62	0.850	0.027	0.299	-0.046
63	0.769	-0.054	0.246	-0.099
64	0.723	-0.100	0.132	-0.213
65	0.647	-0.176	0.172	-0.172
66	0.935	0.112	0.394	0.050
67	0.613	-0.211	0.192	-0.153
68	0.760	-0.063	0.157	-0.188
Mean	0.8230	0.0000	0.3447	0.0000
Variance	0.0849	0.0849	0.0338	0.0338
Std. Dev.	0.2914	0.2914	0.1837	0.1837

Now consider a graph of these two variables. Let’s call the income variable X1, and the ownership variable X2. A glance at the scatter of points suggests a fairly strong, although far from perfect, correlation. This makes sense: on average, higher-income households are much more likely to be able to afford to own their own homes, and, conversely, many of the circumstances that allow people to gain access to homeownership also help them as they try to earn income and build wealth. This is not a perfect relationship – some neighborhoods have high-income residents who choose to rent, while in other places we might find comparatively low-income residents who make substantial sacrifices to achieve ownership. But the relationship is still quite strong, and so it suggests that we might be able to describe important aspects of these neighborhoods if we were to *combine* the information in the two variables. Income and homeownership seem to be capturing different facets of the same thing, and so it would be valuable to have a systematic way of distilling these two measures into a composite measure.

An obvious first step is just to glance at the graph, and see how the upward slope of the scatter of points suggests a separate axis, somewhere between our measure of income (X1) and ownership (X2), that would capture more of what's going on than either of the original variables alone. We could just look at the graph and sketch in an approximation. Suppose we put in a new axis, which we'll call X*, that seems to capture the general drift of the points in the graph.

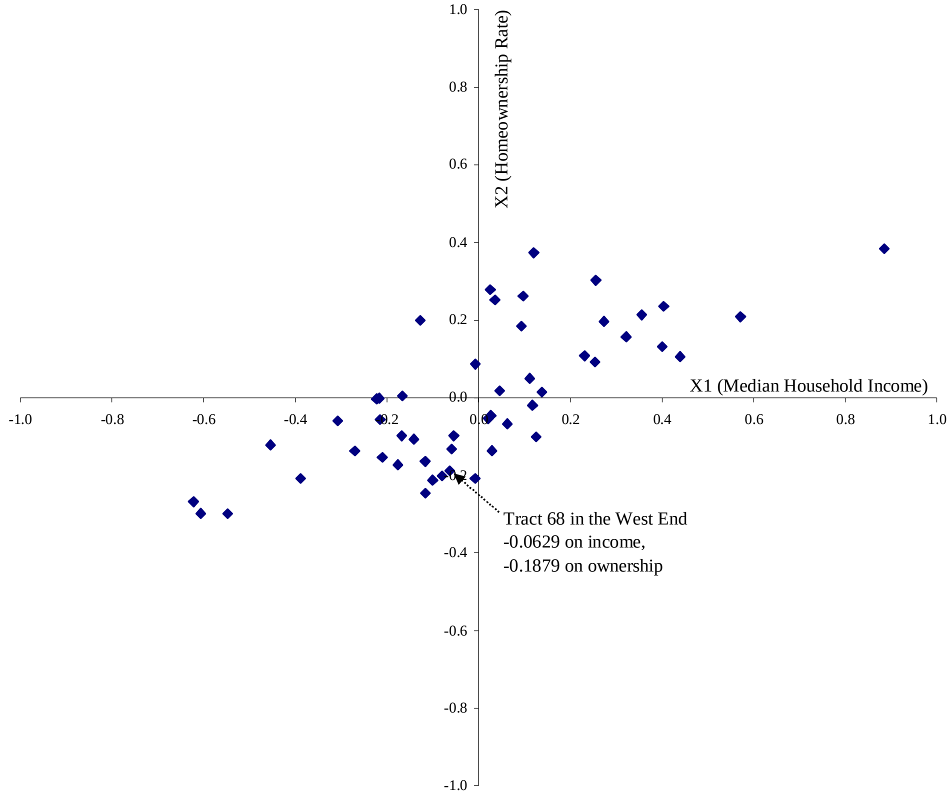

Figure 3. Graph of Median Household Income and Homeownership Rate for Census Tracts in Central Vancouver, 2001. Data Source: Statistics Canada (2003). *Electronic Profiles, Census Metropolitan Areas, Tracted Census Agglomerations, and Census Tracts, 2001 Census.* Release 95F0495XCB2001005. Ottawa: Statistics Canada.

The principles of geometry come in handy at this point, because the relationships in this graph follow all of the rules of right-hand triangles. The axis we've sketched in forms an angle with the original variable X1, and this angle (let's call it theta, θ) allows us to use the formulas for the sine, cosine, and tangent of a right-angle triangle to work out several key relationships. Specifically, these principles of geometry allow us to project all the observations onto the new axis. The perpendicular projection of any point on the new axis will intersect at a point whose distance from the origin can be expressed as x* = (cos θ × x1) + (sin θ × x2).[13]

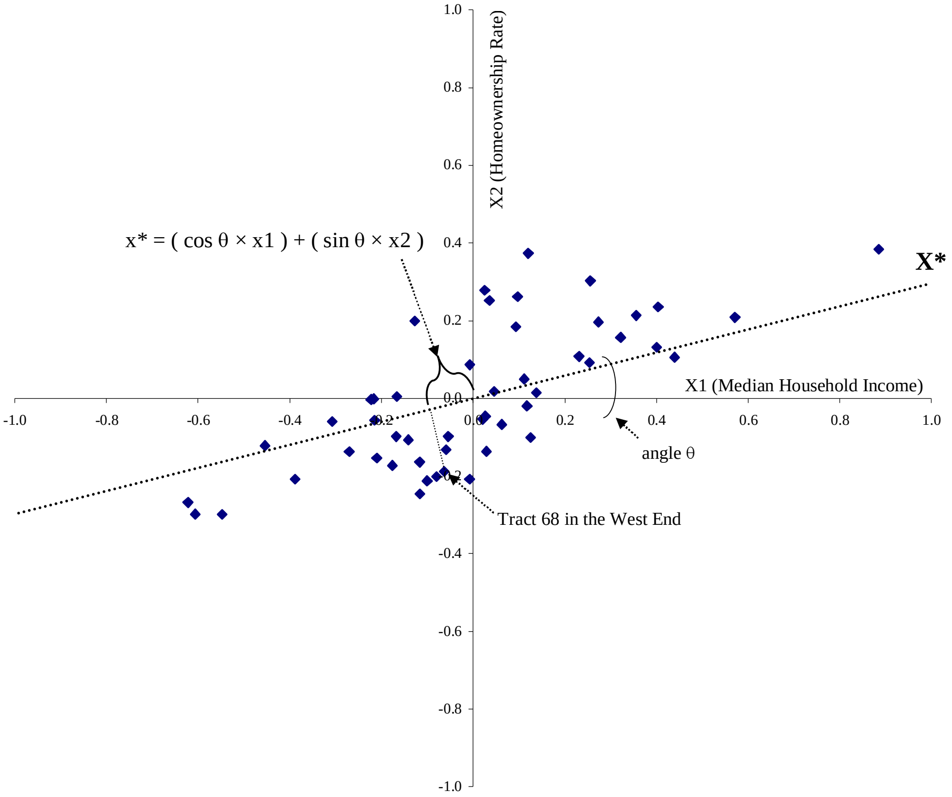

Figure 4. Projecting a Point onto a New Composite Variable.

So if we created an axis at an angle of 20 degrees and we wished to project the point for Census Tract 68 in the West End, the calculation would be

x* = (cos θ × x1) + (sin θ × x2)
x* = (cos (20) × -0.0629) + (sin (20) × -0.1879)
x* = (0.9396 x -0.0629) + (0.3420 × -0.1879)
x* = (-0.0591 + -0.0643)
x* = -0.1234

For any given value of theta, then, it is a simple matter to work out the values of X* for all of our fifty neighborhoods of Central Vancouver. When θ is 20 degrees, the calculations are as shown in Table 2.

[13] For the derivation of this and similar equations, see an intermediate mathematics or geometry text. There is also a short summary in Chapter 2 of Sharma, Subhash (1996). *Applied Multivariate Techniques.* New York: John Wiley and Sons.

Table 2. Calculating a New Variable (X*) when Angle Theta is 20.°

Census	Income	Ownership	
Tract	X1	X2	X*
38	-0.1681	-0.0986	-0.1917
39.01	-0.0589	-0.1320	-0.1005
39.02	0.0456	0.0181	0.0490
40.01	-0.0077	-0.2081	-0.0784
40.02	0.0294	-0.1364	-0.0190
41.01	0.0631	-0.0671	0.0363
41.02	0.3557	0.2133	0.4072
42	0.2735	0.1966	0.3243
43.01	0.5715	0.2083	0.6082
43.02	0.8852	0.3837	0.9631
44	0.4037	0.2359	0.4601
45.01	0.4390	0.1060	0.4488
45.02	0.2541	0.0916	0.2701
46	0.0213	-0.0533	0.0018
47.01	0.1176	-0.0193	0.1039
47.02	0.1254	-0.1012	0.0832
48	0.1382	0.0145	0.1348
49.01	0.3216	0.1570	0.3559
49.02	0.2311	0.1081	0.2541
50.02	-0.2703	-0.1369	-0.3009
50.03	-0.2167	-0.0015	-0.2042
50.04	-0.3071	-0.0597	-0.3090
51	0.0253	0.2786	0.1190
52.01	-0.1273	0.1986	-0.0517
52.02	0.0355	0.2518	0.1194
53.01	0.0969	0.2616	0.1805
53.02	0.2557	0.3023	0.3437
54.01	0.0932	0.1845	0.1507
54.02	0.1206	0.3736	0.2411
55.01	-0.2228	-0.0029	-0.2103
55.02	-0.1660	0.0051	-0.1543
56.01	-0.3882	-0.2080	-0.4359
56.02	-0.1409	-0.1067	-0.1689
57.01	-0.5476	-0.2994	-0.6170
57.02	-0.4540	-0.1218	-0.4683
58	-0.6065	-0.2980	-0.6718
59.03	0.4006	0.1321	0.4216
59.04	-0.2146	-0.0563	-0.2209
59.05	-0.0070	0.0864	0.0230
59.06	-0.6221	-0.2682	-0.6763
60.01	-0.0800	-0.2019	-0.1443
60.02	-0.1165	-0.1640	-0.1656
61	-0.1161	-0.2456	-0.1931
62	0.0268	-0.0456	0.0096
63	-0.0545	-0.0986	-0.0849
64	-0.1002	-0.2126	-0.1669
65	-0.1761	-0.1724	-0.2244
66	0.1120	0.0498	0.1223
67	-0.2105	-0.1531	-0.2502
68	-0.0629	-0.1879	-0.1234
Variance	0.0849	0.0338	0.1044

Now look at the line at the bottom of this table, where I've calculated the variance for each of our variables. Recall that variance is simply a measure of how how much a set of values vary from the mean. Our income variable has a total variance of 0.0849, while the ownership rate has a variance of 0.0338. But the new composite variable has a variance of 0.1044. This is 87.96 percent of the total variance in the dataset. In other words, the new composite variable captures more information about our neighborhoods of Central Vancouver than either income or homeownership alone.

Each value of theta will yield a different set of scores on X*, and will also result in distinct values for the total variance term. If we calculate all of these values for different values of theta, we can compare the variance of the new axis to the total for our dataset (Table 3). Note that as we increase the angle, the new variable accounts for an increasing fraction of total variance, until we reach some point – in this case, it's 28.55 degrees -- and then declines; by the time theta is 90 degrees, the new axis is equivalent to X2, and not surprisingly, its total variance is 0.0338, exactly the same as variable X2. Figure 5 shows a graph of the total proportion of variance accounted for by the new component X* for different values of angle theta.

Table 3. Variance of X* for Different Angles.

Angle	Variance	Proportion
0	0.0849	0.7156
10	0.0969	0.8167
20	0.1044	0.8796
28.55	0.1065	0.8973
30	0.1064	0.8968
40	0.1028	0.8661
50	0.0939	0.7912
60	0.0808	0.6812
70	0.0652	0.5494
80	0.0488	0.4116
90	0.0338	0.2844

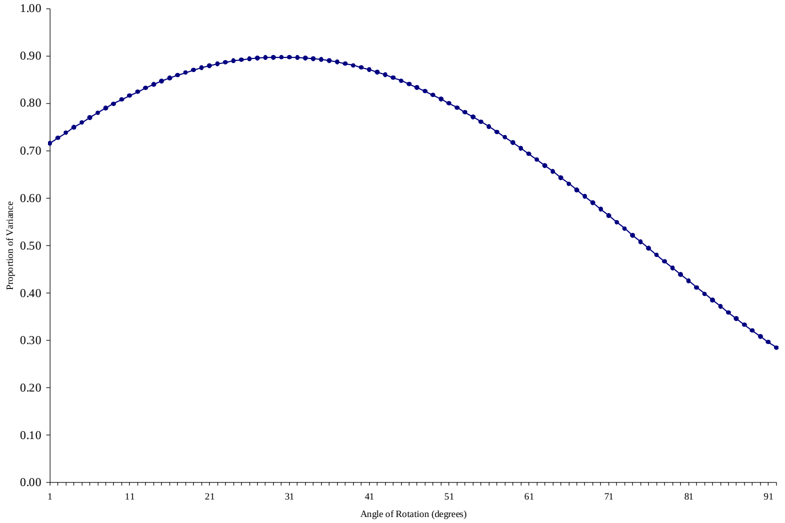

Figure 5. Total Variance of X* for different Values of Theta.

The purpose of principal components analysis is to define X* in such a way as to account for the largest possible proportion of total variance. There is one, *and only one*, angle at which the new, composite variable will account for the maximum proportion of the information contained in the original variables. For this small dataset describing income and homeownership in Central Vancouver, that angle is 28.55 degrees; at this point, the first component, X*, accounts for 89.73 percent of the total information included in the two separate variables. Still, what about the remaining 10.27 percent? To account for this, we lay in another axis, perpendicular to X* (see Figure 6). If we call this second new variable X**, a similar set of geometric right-angle principles apply, and we can project the observations onto this new axis with another equation:
$x^{**} = (-\sin \theta \times x1) + (\cos \theta \times x2)$

When theta is 28.55 degrees, the calculation for Tract 68 in the West End is:
$x^{**} = (-\sin (28.55) \times -0.0629) + (\cos (28.55) \times -0.1879)$
$x^{**} = (-0.4799 \text{ x } -0.0629) + (0.8784 \times -0.1879)$
$x^{**} = (0.0301 + -0.1650)$
$x^{**} = -0.1349$

Table 4 shows the calculation of these values for all of the census tracts in Central Vancouver.[14]

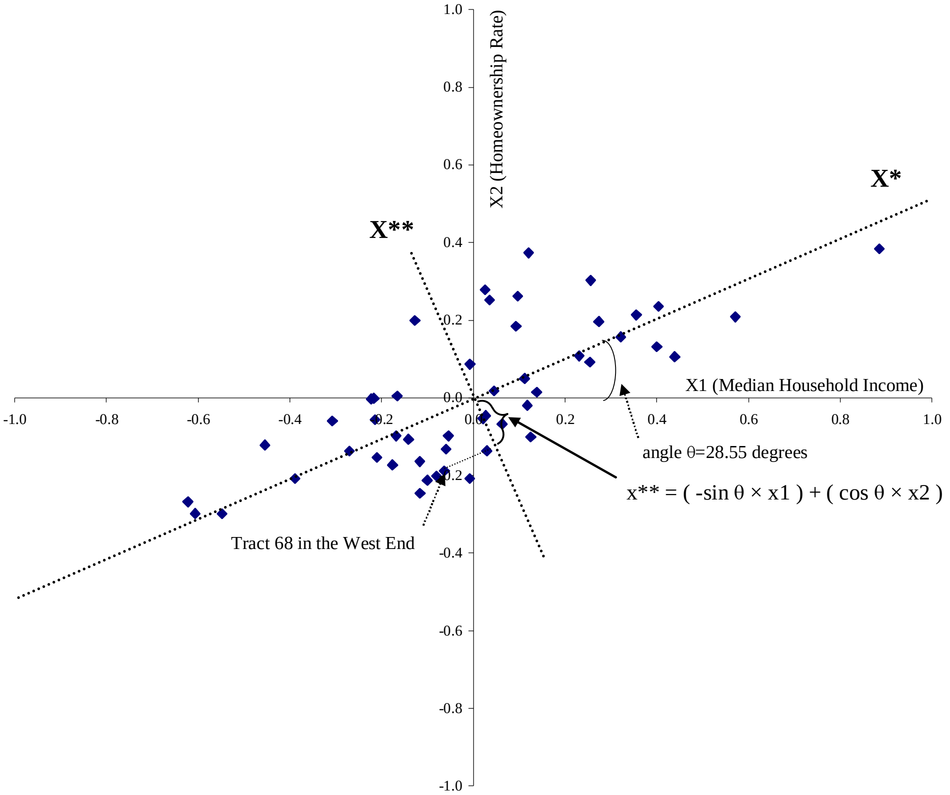

Figure 6. Components X* and X** when Theta is 28.55 Degrees.

[14] The calculations shown above involve a slight rounding error, since I only took the numbers out to four decimal places. So the value for x** above is -0.1349, versus -0.1350 in Table 4.

Table 4. Calculating New Variables X* and X** Angle Theta is 28.55 Degrees.

Census Tract	X1	X2	X*	X**
38	-0.1681	-0.0986	-0.1948	-0.0063
39.01	-0.0589	-0.1320	-0.1148	-0.0878
39.02	0.0456	0.0181	0.0487	-0.0059
40.01	-0.0077	-0.2081	-0.1062	-0.1791
40.02	0.0294	-0.1364	-0.0393	-0.1339
41.01	0.0631	-0.0671	0.0233	-0.0891
41.02	0.3557	0.2133	0.4144	0.0174
42	0.2735	0.1966	0.3342	0.0420
43.01	0.5715	0.2083	0.6015	-0.0902
43.02	0.8852	0.3837	0.9610	-0.0860
44	0.4037	0.2359	0.4674	0.0142
45.01	0.4390	0.1060	0.4363	-0.1167
45.02	0.2541	0.0916	0.2670	-0.0410
46	0.0213	-0.0533	-0.0068	-0.0570
47.01	0.1176	-0.0193	0.0941	-0.0732
47.02	0.1254	-0.1012	0.0618	-0.1488
48	0.1382	0.0145	0.1283	-0.0533
49.01	0.3216	0.1570	0.3575	-0.0158
49.02	0.2311	0.1081	0.2546	-0.0155
50.02	-0.2703	-0.1369	-0.3029	0.0089
50.03	-0.2167	-0.0015	-0.1911	0.1023
50.04	-0.3071	-0.0597	-0.2983	0.0943
51	0.0253	0.2786	0.1554	0.2327
52.01	-0.1273	0.1986	-0.0169	0.2353
52.02	0.0355	0.2518	0.1515	0.2042
53.01	0.0969	0.2616	0.2101	0.1834
53.02	0.2557	0.3023	0.3691	0.1434
54.01	0.0932	0.1845	0.1700	0.1175
54.02	0.1206	0.3736	0.2845	0.2705
55.01	-0.2228	-0.0029	-0.1971	0.1039
55.02	-0.1660	0.0051	-0.1434	0.0838
56.01	-0.3882	-0.2080	-0.4404	0.0029
56.02	-0.1409	-0.1067	-0.1748	-0.0264
57.01	-0.5476	-0.2994	-0.6241	-0.0013
57.02	-0.4540	-0.1218	-0.4570	0.1100
58	-0.6065	-0.2980	-0.6751	0.0281
59.03	0.4006	0.1321	0.4150	-0.0755
59.04	-0.2146	-0.0563	-0.2154	0.0531
59.05	-0.0070	0.0864	0.0352	0.0792
59.06	-0.6221	-0.2682	-0.6746	0.0618
60.01	-0.0800	-0.2019	-0.1668	-0.1391
60.02	-0.1165	-0.1640	-0.1807	-0.0884
61	-0.1161	-0.2456	-0.2194	-0.1603
62	0.0268	-0.0456	0.0018	-0.0529
63	-0.0545	-0.0986	-0.0950	-0.0606
64	-0.1002	-0.2126	-0.1896	-0.1389
65	-0.1761	-0.1724	-0.2371	-0.0673
66	0.1120	0.0498	0.1222	-0.0098
67	-0.2105	-0.1531	-0.2581	-0.0339
68	-0.0629	-0.1879	-0.1451	-0.1350
Variance	0.08492	0.03376	0.1065	0.0122

Note that this time, the variance term (0.0122) is much smaller than that of the each of the original variables. But this is 10.27 percent of the total variance, exactly the amount “left over” after we defined the first component. The variances of X* and X** sum to the same value as the original variables. The first principal component, X*, captures the vast majority of the total information that we began with, but the second picks up the rest. This is the basic approach of principal components analysis: performing simple combinations of variables into new axes, to create new, composite variables that convey the information in the original dataset in new ways. The results of a principal components analysis have several valuable properties.

1. The maximum number of new components is always equal to the number of original variables. In this simple example, we only used two variables, so our analysis yields two components; but there is nothing keeping us in the boring realm of two or even three dimensions. The number of variables we can use is virtually unlimited.

2. The cumulative variance of all the new *components* will always sum to the same value as the total variance of the original *variables* that we began with.

3. The first component will always account for the largest possible share of the total information in the dataset. The second component will account for the second-largest share of total variance, the third component will account for the third-largest share, and so on.

4. The units in which the variables are measured will affect the results of a principal components analysis: if one of the variables has a very large variance, then it will tend to dominate the results. Therefore, if you're working with measures that have wildly different scales (say, percentages versus many thousands of dollars for income), it's a good idea to standardize things first (expressing each observation in terms of its standard deviation from the mean for each variable).

5. The components can be derived geometrically, as in this example, but they can also be obtained through matrix algebra. The variance accounted for by each component is often called the **eigenvalue** for that component, because it is equal to the first latent root of the correlation matrix; 'eigen' is German for 'root.'

6. The relationship between the original variables and the components is summarized in an **eigenvector**. The elements of the vector are called ***loadings***, and they measure the strength of the relationship between the original variable and the new components. Loadings range from -1.0 to +1.0. Loadings that are more positive or more negative indicate a stronger relationship between the original variable and the new component; loadings that are closer to zero indicate a weak relationship. The *squared* loading measures the proportion of the variance of the original variable that is captured by the new component. The sum of the squared loadings for a particular variable, across multiple components, is known as the ***communality*** of the variable: the proportion of the total information in the original variable that is captured by all of the components taken together.

7. The new components are orthogonal (at right angles) and are therefore independent and uncorrelated.

8. The coordinates of observations on the new axes are known as the component scores. These are the values shown in the X* and X** columns above in Table 4.

All these calculations might seem rather tedious. But once we understand what a principal component is by visualizing it in these geometric terms, then it’s a simple matter to do the calculations quickly in computer-based statistical software. The following code reads the data shown in Table 1, and performs a simple principal components analysis.

```
libname g350 "c:\sasdat\g350";
data g350.vsimple(compress=yes);
            infile "c:\sasdat\g350\v_simple.csv" delimiter="," missover;
            input tract mhhinc hown;
            run;
proc princomp data=g350.vsimple cov out=stemp;
            var mhhinc hown;
            title 'simple pca illustration';
            run;
proc print; id tract; var prin1 prin2 mhhinc hown; run;
```

This gives us the following output. The red letters correspond to the descriptions and explanations below.

```
                                      simple pca illustration     06:30 Monday, October 22, 2007  12

                                       The PRINCOMP Procedure

                                      Observations          50
                                      Variables              2

                                         Simple Statistics

                                              mhhinc              hown

                              Mean      0.8230160192      0.3447149355
                              StD       0.2914144461      0.1837272437

                                         Covariance Matrix

                                               mhhinc              hown

                             mhhinc      0.0849223794      0.0396031356
                             hown        0.0396031356      0.0337557001


                                   Total Variance    0.1186780795
1
                                Eigenvalues of the Covariance Matrix

                            Eigenvalue    Difference    Proportion    Cumulative
2
                       1    0.10648685    0.09429561        0.8973        0.8973
                       2    0.01219123                      0.1027        1.0000

                                            Eigenvectors

                                                Prin1         Prin2
3
                                 mhhinc      0.878243      -.478215
                                 hown        0.478215      0.878243

                        tract      Prin1       Prin2      mhhinc      hown
4

                        38.00    -0.19476    -0.00622    0.65495    0.24612
                        39.01    -0.11486    -0.08778    0.76412    0.21270
                        39.02     0.04868    -0.00593    0.86860    0.36278
                        40.01    -0.10624    -0.17906    0.81534    0.13665
                        40.02    -0.03939    -0.13384    0.85242    0.20833
```

```
41.01     0.02328    -0.08913    0.88608    0.27757
41.02     0.41443     0.01726    1.17873    0.55806
42.00     0.33423     0.04189    1.09652    0.54134
43.01     0.60147    -0.09037    1.39447    0.55298
43.02     0.96095    -0.08633    1.70825    0.72843
44.00     0.46737     0.01406    1.22675    0.58057
45.01     0.43627    -0.11687    1.26205    0.45070
45.02     0.26699    -0.04110    1.07715    0.43629
46.00    -0.00681    -0.05701    0.84429    0.29139
47.01     0.09405    -0.07323    0.94063    0.32538
47.02     0.06175    -0.14882    0.94842    0.24354
48.00     0.12831    -0.05338    0.96123    0.35920
49.01     0.35749    -0.01593    1.14459    0.50168
49.02     0.25464    -0.01555    1.05408    0.45283
50.02    -0.30291     0.00901    0.55268    0.20777
50.03    -0.19108     0.10233    0.60627    0.34321
50.04    -0.29827     0.09439    0.51592    0.28497
51.00     0.15544     0.23260    0.84830    0.62333
52.01    -0.01681     0.23528    0.69573    0.54331
52.02     0.15154     0.20417    0.85847    0.59649
53.01     0.21019     0.18337    0.91992    0.60628
53.02     0.36917     0.14324    1.07873    0.64706
54.01     0.17007     0.11748    0.91620    0.52922
54.02     0.28455     0.27045    0.94359    0.71831
55.01    -0.19705     0.10401    0.60022    0.34183
55.02    -0.14338     0.08386    0.65699    0.34979
56.01    -0.44040     0.00301    0.43480    0.13675
56.02    -0.17481    -0.02633    0.68208    0.23799
57.01    -0.62412    -0.00106    0.27539    0.04532
57.02    -0.45698     0.11015    0.36900    0.22292
58.00    -0.67513     0.02832    0.21654    0.04673
59.03     0.41497    -0.07559    1.22361    0.47678
59.04    -0.21540     0.05314    0.60843    0.28837
59.05     0.03518     0.07920    0.81604    0.43110
59.06    -0.67462     0.06201    0.20088    0.07656
60.01    -0.16683    -0.13900    0.74297    0.14286
60.02    -0.18074    -0.08829    0.70651    0.18074
61.00    -0.21941    -0.16023    0.70695    0.09907
62.00     0.00177    -0.05287    0.84986    0.29913
63.00    -0.09499    -0.06054    0.76854    0.24612
64.00    -0.18966    -0.13885    0.72285    0.13208
65.00    -0.23710    -0.06723    0.64694    0.17228
66.00     0.12219    -0.00985    0.93504    0.39450
67.00    -0.25811    -0.03383    0.61252    0.19157
68.00    -0.14510    -0.13499    0.76013    0.15677
```

Output 1. PCA Analysis for the Simple Illustration.

This kind of statistical output would look a bit obtuse or intimidating at first, if we had not first tried to visualize what the computer is doing. If you look at things carefully it makes a lot of sense. Just keep in your mind's eye those simple graphs of neighborhoods, and the angles between the components and the original variables.

1. First, note that the software provides a few basic statistics for the two variables, mhhinc (the median household income ratio) and hown (the homeownership rate). One of the items reported is the **total variance** – the total amount of information conveyed by the differences between the neighborhoods, measured on these two simple variables, for part of Central Vancouver.

2. Second, note the **eigenvalues**: the first one captures 89.73 percent of the total variance in the dataset, and the second captures the remaining 10.27 percent. Note that the eigenvalues here match the "variance" row for X* and X** at the bottom of Table 4.

3. Third, consider the **eigenvectors**. These show the relations between the original variables and the new components (which the software calls Prin1 and Prin2). The household income ratio shows a loading of 0.8782 onto the first component. This is the cosine of the angle between the income variable and the first component. The squared loading (0.878243 * 0.878243 = 0.7713)

indicates that the first component captures more than three-quarters of the information conveyed in the income variable. The second component captures the remaining 22.87 percent.

4. Fourth, the printout includes the values for the original measures of income and homeownership, as well as the **component scores** – the values of each neighborhood projected onto these new composite components, which describe a *combination of income and homeownership.* At this point, we could map all of the neighborhoods of this part of the city on these component scores: the map would show the spatial distribution of the *conceptual relations between income and homeownership,* and thus would capture more information than either of these variables alone.

How to Interpret a "Classical" Factor Analysis for Vancouver.

We could, of course, spend more time examining homeownership and income in this part of Central Vancouver. Still, aren't we getting just a little bit bored with just these two variables, which seem to be telling us very similar things? Stretch your imagination from the sterile, two-dimensional graph shown above, and visualize in you're mind's eye multidimensional world in which we can portray many different aspects of each neighborhood in the metropolitan region. First, let's consider an update in the style of the classical versions of social area analysis, which proposed that the most important contrasts could be captured in economic status, family status, and ethnic status. I selected about a dozen and a half basic measures of the social characteristics for each of the census tracts in the Vancouver metropolitan area. Specific variables fall into three categories:

1. Economic and class-related variables: percentage of persons in households with income below the 'low-income cutoff'; median household income as share of metropolitan average; share of households with annual income of more than $100,000; average dwelling value as share of metropolitan average; homeownership rate; and unemployment rate.

2. Family and demographic characteristics: percentage of persons over age 15 who are single, divorced, and separated; percentage of families who are married couples with children, and who are female lone parents.

3. Ethnic diversity characteristics: percentage of persons identifying themselves as visible minorities, as Chinese, South Asian, Black, Filipino, or South Asian.

I then performed a **factor analysis** of these variables; think of principal components analysis as a simple, special kind of factor analysis. There are some additional details and complications associated with shifting from principal components analysis to factor analysis. But these tedious details are more important for statisticians than urban geographers. For our purposes, the interpretation of the results of a factor analysis is quite similar to the simple principal components analysis described above.

Output 2, below, provides some of the statistical results from the factor analysis. The red letters correspond to the descriptions and explanations on the following pages.

```
                         Classical Social Area Analysis Results                         57
                                                        08:18 Wednesday, October 31, 2007

                                    The FACTOR Procedure
                              Initial Factor Method: Principal Components

                                 Prior Communality Estimates: ONE

                      Eigenvalues of the Correlation Matrix: Total = 17  Average = 1

                              Eigenvalue    Difference    Proportion    Cumulative
1

                         1    6.76551233    4.12627613        0.3980        0.3980
                         2    2.63923620    0.85916902        0.1552        0.5532
                         3    1.78006718    0.71302944        0.1047        0.6579
                         4    1.06703774    0.23908874        0.0628        0.7207
                         5    0.82794900    0.02010359        0.0487        0.7694
                         6    0.80784541    0.08095549        0.0475        0.8169
                         7    0.72688992    0.09808125        0.0428        0.8597
                         8    0.62880867    0.17389886        0.0370        0.8967
                         9    0.45490981    0.05683560        0.0268        0.9234
                        10    0.39807421    0.12283218        0.0234        0.9468
                        11    0.27524203    0.08903359        0.0162        0.9630
                        12    0.18620844    0.04203474        0.0110        0.9740
                        13    0.14417370    0.02420159        0.0085        0.9825
                        14    0.11997211    0.03827526        0.0071        0.9895
                        15    0.08169685    0.01727275        0.0048        0.9943
                        16    0.06442410    0.03247182        0.0038        0.9981
                        17    0.03195228                      0.0019        1.0000

                          4 factors will be retained by the NFACTOR criterion.

                                       Rotated Factor Pattern

                                                    Factor1       Factor2       Factor3       Factor4
2
 pchange    percentage population change           -0.06783       0.01869       0.00702       0.89937
 lico       total incidence of low income           0.78676      -0.16803      -0.44435       0.08658
 mhhinc     median household income ratio          -0.60772       0.47703       0.50482      -0.02895
 elite      share of households over 100k          -0.53918       0.63710       0.35014      -0.02112
 avval      average dwelling value ratio           -0.19024       0.80369       0.06886      -0.12580
 hown       homeownership rate                     -0.45626       0.30402       0.75467      -0.04124
 unemp      unemployment rate                       0.69665      -0.19445      -0.26383       0.10865
 single     single population share                 0.30792       0.00365      -0.78003       0.03223
 divorce    divorced population share              -0.14347      -0.58846      -0.71331      -0.13556
 separ      separated population share              0.07701      -0.81772      -0.37413      -0.11085
 mckids     married couples with children           0.10997       0.38841       0.83008       0.00686
 flone      female lone parent households           0.53794      -0.52463      -0.17870      -0.27697
 vm_ch      vismin Chinese                          0.68163       0.57842      -0.02113      -0.02139
 vm_sa      vismin South Asian                      0.23125      -0.31673       0.53626       0.39113
 vm_bl      vismin Black                            0.29046      -0.51110      -0.04005       0.04106
 vm_fl      vismin Filipino                         0.66557      -0.20752       0.11538      -0.03273
 vm_se      vismin Southeast Asian                  0.71855      -0.08820       0.09570      -0.08367

                             Final Communality Estimates: Total = 12.251853
3
       pchange           lico          mhhinc           elite            avval            hown

    0.81386531      0.85217785      0.85255467      0.81966152      0.70268184      0.87182447

         unemp          single         divorce           separ           mckids           flone

    0.60453789      0.70431308      0.89406302      0.82685650      0.85202855      0.67326766

                 vm_ch           vm_sa           vm_bl           vm_fl           vm_se

            0.80009308      0.59435685      0.34888652      0.50042668      0.54025797
```

Output 2. Factor Analysis for a Classical Social Area Analysis of Vancouver.

In the next few pages, we will consider three broad sets of questions that this kind of analysis helps us to explore. I'll show you how to evaluate and interpret the results of this analysis, which is based on the comparatively simple, classical approach to social area analysis that was first devised by Shevky and Bell more than half a century ago. After you consider how to interpret this simple example, I'll give you a much more interesting set of results that capture a few more of the contemporary nuances of urban social patterns as described by Murdie and Teixeira and Knox and McCarthy.

1. *Are there any general trends amidst the complexity of the urban social mosaic?*

Our first set of questions is simple: is it possible to distill the remarkable diversity of the Vancouver social fabric, as measured in that dozen-and-a-half variables, into a smaller number of composite dimensions of urban social space? The entire history of social area analysis and factorial ecology would lead us to suspect that the answer is yes. Our results confirm these expectations: the first eigenvalue accounts for 39.8 percent of all of the variance in our original set of seventeen variables.[15] The second accounts for 15.5 percent, the third for 10.5 percent, and the fourth for 6.3 percent. The cumulative percentage of all information captured in the first four components is just over 72 percent: we can account for about seven-tenths of all of the information in our chosen set of social indicators if we distill them into four main composite variables or factors. Many of the varied and specific measures for neighborhoods seem to be capturing different aspects of the same underlying general conditions.

2. *What are the most important dimensions of the urban mosaic?*

But what are these dimensions? We can explore this question by studying the "rotated factor pattern," which presents the loadings for each of the original measure on the composite, underlying factor. The squared value of each loading tells us the proportion of variance in the raw indicator that is captured by the new, composite variable. Consider the value for lico, the share of persons in households with incomes below the census-defined "low-income cutoff"; this variable shows a loading of about 0.79 with Factor 1. The square of this value (about 0.62) indicates that 62 percent of the information in the original variable, which measures the prevalence of low-income households across the many neighborhoods of Vancouver and its suburbs, can be captured in the new, composite factor.

We can interpret the meaning of this composite factor by considering the individual variables that show very strong positive *or negative* loadings.[16] Factor 1 *distinguishes and separates* neighborhoods that have large shares of low-income people (+0.79) and high rates of unemployment (+0.70), on the one hand, from neighborhoods that have higher median household incomes (-0.61), and shares of households with incomes over $100,000 per year (-0.54). This pattern is generally consistent with the simple notion of economic status in social area analysis. But other complications are apparent: variables for the proportion of persons identifying themselves as Chinese, Filipino, and Southeast Asian also post high loadings on Factor 1. This

[15] This compares with 5.88 percent for each individual variable. In this analysis we're working with standardized data, so each individual "raw" variable accounts for 1/17 (5.88 percent) of the entire variance. But Factor 1, accounts for 39.8 percent – making it 6.77 times more efficient as a way of capturing information than any of the original variables.

[16] Use judgment and subjectivity in deciding how large a value must be to qualify as a "very strong positive" or "very strong negative" loading. I recommend ignoring *many* of the loadings when they begin to dip below 0.50. There are exceptions to this general guidance, however, especially if you find several variables all posting values just shy of the threshold.

suggests the classical concepts of economic status and ethnic status – usually portrayed as aspects of social structure that could be disentangled into independent, separate dimensions – are intextricably woven together in contemporary Vancouver. In general, the interrelations suggest that Factor 1 is capturing something that we might call "Class and Racial-Ethnic Inequality." Factor 2, by contrast, captures important aspects of the diversity of household and family circumstances in different neighborhoods. We see strongly negative loadings for divorcees, separated couples, and women raising children alone, and strongly positive loadings for house prices, high-income households, and median household income. This dimension of urban social structure seems to reflect the severe challenges associated with family breakups, which often make things difficult for divorced or separated lone mothers unless they move to neighborhoods with a sufficient supply of affordable housing. But these neighborhoods are moderate-income areas, and not the poorest districts: note the very weak loading for the low-income cutoff variable (-0.17). And as in the case of the first factor, family and demographic diversity is intertwined with certain racial and ethnic contrasts: the affordable neighborhoods where divorcees and single mothers are living tend to have fewer people identifying themselves as Chinese, and proportionally more people identifying themselves as Black.[17] Overall, though, the pattern of loadings seem to merit calling this factor "Family Breakup."

A third dimension of urban social structure hints at other facets of "family-oriented" neighborhoods across the metropolis. This composite measure correlates strongly with high rates of homeownership (+0.75) married couples with children (+0.83), and median household income (+0.50) and shows opposite, negative correlations with single, unmarried persons (-0.78) and divorcees (-0.71). We might call this an axis of "Traditional Families." There is also a significant positive loading for persons identifying themselves as South Asian, repeating once again the pattern of interwoven family-status and ethnic-status patterns. These results – the overall configuration of loadings for both Factors 2 and 3 – give us exactly the kind of evidence we need to understand what Murdie and Teixeira mean when they describe the "increased fragmentation of the family status factor" relating to the sweeping "changes in family and household structure" in postindustrial Canadian society.[18]

A fourth dimension of urban social structure is simple: only one variable loads significantly onto this factor, and it's a strong positive. Percentage population change posts a loading of 0.89 on Factor 4, clearly justifying that we label this dimension "Growth."

It can sometimes make your eyes glaze over to try to sift through detailed tables of loadings, eigenvalues, and the like. But the payoff comes when you sort through the inter-relations between all the different ways of looking at social diversity in the metropolis, you develop hunches of what these new composite variables or "factors" are really describing as they distill the various indicators, and then you map the factor scores for each observation. Figure 7 shows the tract scores for Factor 1, which we suggested above is reflecting the broad contours of class and racial inequality in the region. This is the dominant dimension of socio-spatial structure: it

[17] The loading for Chinese people is positive, and the loading for Blacks is negative. But what matters in this interpretation is that the variable for Blacks is loading in the same negative direction for divorcees, separated persons, and female lone parents; while the variable for Chinese is loading in the opposite (positive) direction.
[18] Murdie and Teixeira, "Urban Social Space," p. 159.

captures almost two-fifths of all the information we began with. The darkest purple areas have the lowest scores: given the negative loading for median household income on Factor 1, these purple areas etch out the suburban ring of upper-middle class and elite suburbs, while the strongest positive scores (the dark green neighborhoods) highlight the areas with the greatest concentrations of unemployment and low-income families.

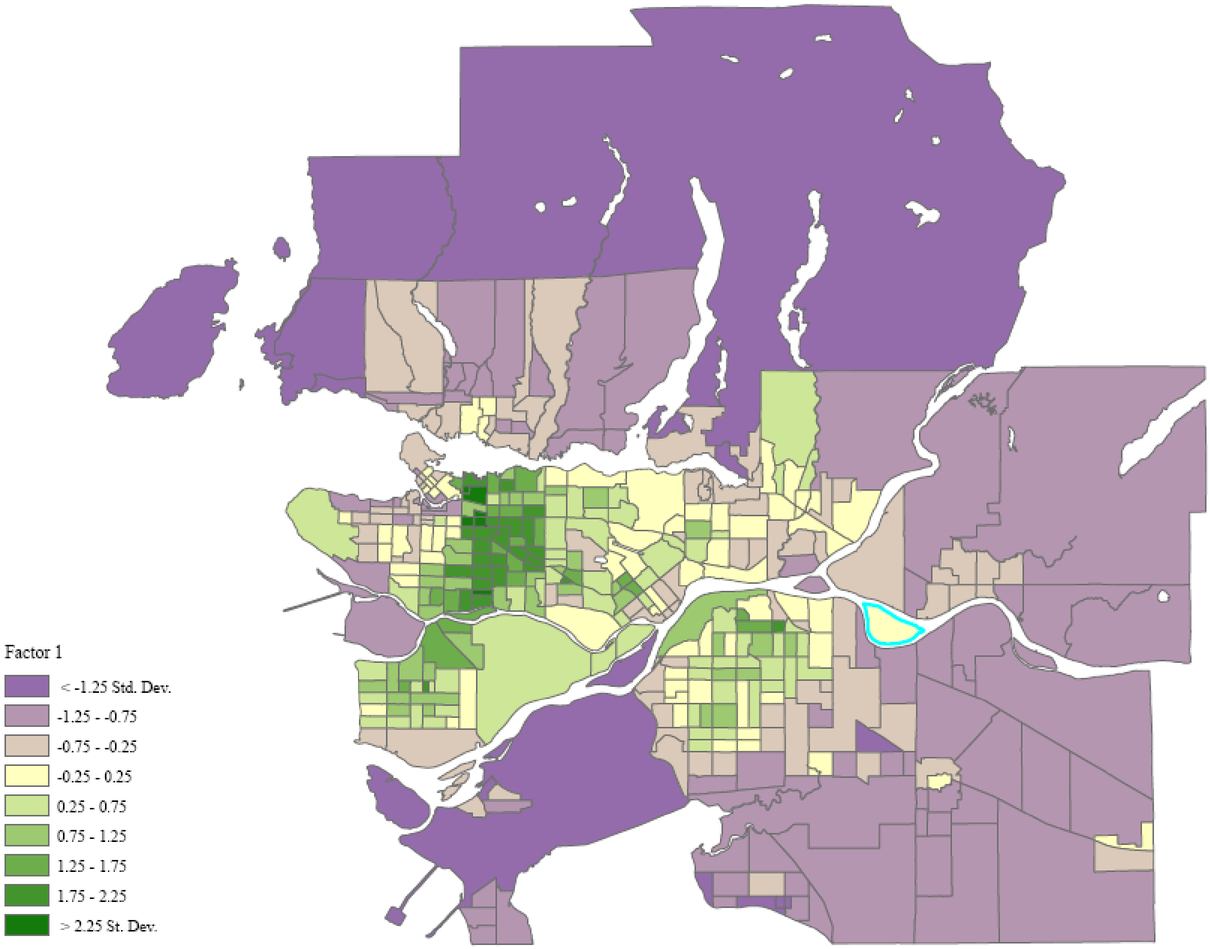

Figure 7. Census Tract Scores for Factor 1 from the Classical Social Area Analysis, Vancouver CMA. Data Source: Statistics Canada (2003). *Electronic Profiles, Census Metropolitan Areas, Tracted Census Agglomerations, and Census Tracts, 2001 Census.* Release 95F0495XCB2001005. Ottawa: Statistics Canada.

The tract scores for Factor 4 reveal an even more vivid spatial partitioning of the metropolis (see Figure 8). This simple ‘Growth’ dimension may not be the most important or dominant aspect of things across the entire metropolis – remember that it only accounts for 6.3 percent of the total information we began with – but it clearly highlights those neighborhoods that are experiencing the most dramatic new development or redevelopment. The darkest green areas have the highest positive factor scores (which, given the positive loading for population change, means that these areas have the highest rates of positive population growth). The map clearly identifies the dramatic restructuring of old office, commercial, and light-manufacturing districts that began in the 1990s with large-scale condo development in Yaletown and other parts of the downtown peninsula, and it also clearly identifies the dramatic expansion of single-family subdivisions in

Surrey and other suburbs. To be sure, we could have simply mapped the single variable of population change, and dispensed with all the jargon about eigenvalues and the like. But the advantage of performing an analysis like this is that the features of components and factors – orthogonal dimensions extracted from the web of inter-related variables – is that we have clear, statistical confirmation that population growth can be identified as *a distinct aspect of urban change that is not inherently bound up with, for example, major cleavages of income inequality or homeownership*. Note that the variable for population change posts a strong loading on Factor 4 only, and posts negligible values on the factors for class and racial-ethnic inequality, family breakup, and traditional families. Our factor analysis demonstrates that the neighborhoods experiencing different rates of population growth include a wide variety of communities in terms of income, ethnic diversity, and family composition: fast-growing neighborhoods include both renter-oriented singles bastions in Yaletown, and broad expanses of Surrey subdivisions with married couples with children. This is what it means to suggest that Factor 4, the ‘Growth’ factor, is independent and orthogonal from all the other measures included in the analysis.

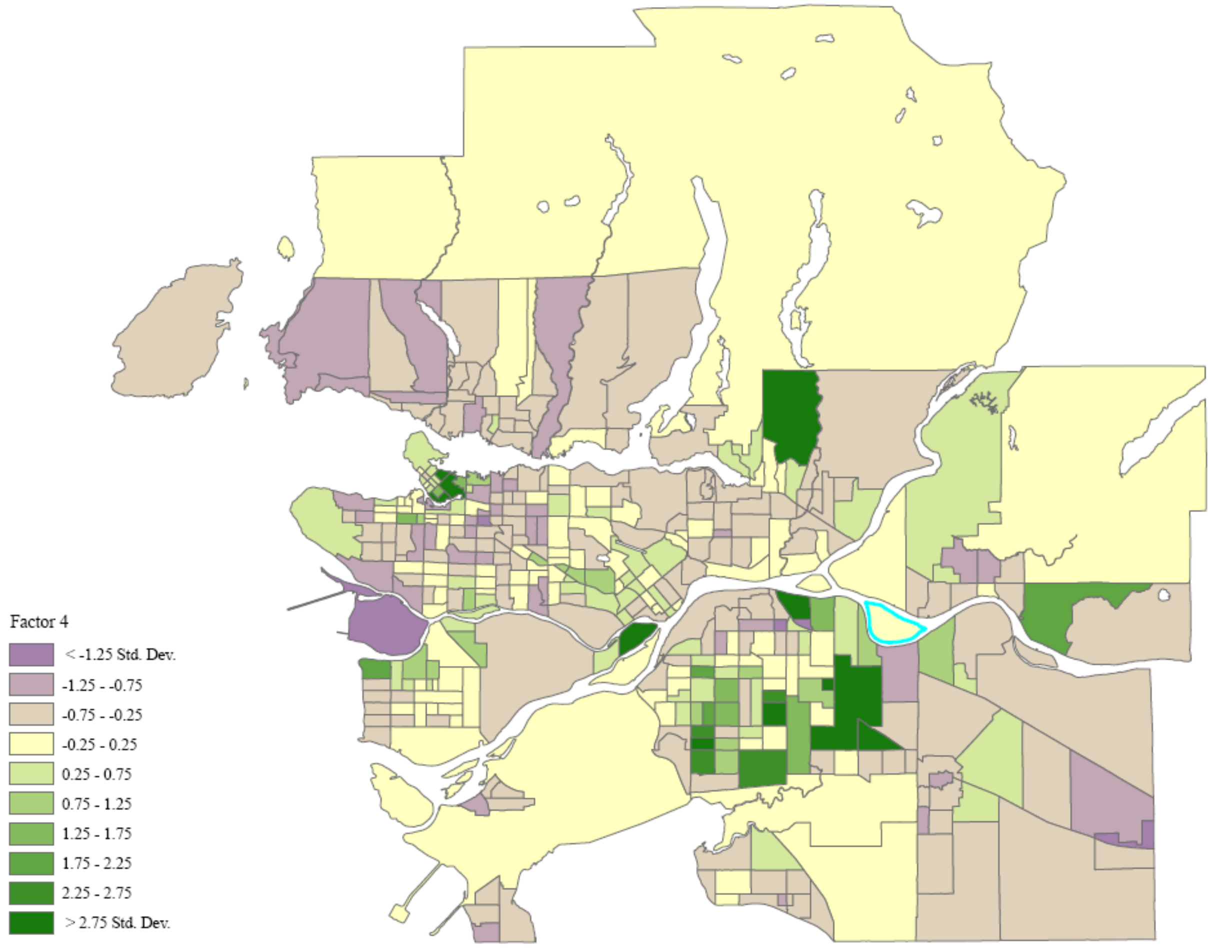

Figure 8. Census Tract Scores for Factor 4 from the Classical Social Area Analysis, Vancouver CMA. Data Source: Statistics Canada (2003). *Electronic Profiles, Census Metropolitan Areas, Tracted Census Agglomerations, and Census Tracts, 2001 Census.* Release 95F0495XCB2001005. Ottawa: Statistics Canada.

Maps of scores for all factors, including Factors 2 and 3, are on the course web site. Studying the geographical pattern of the factor scores, and comparing how certain neighborhoods score across different dimensions, provides a valuable way of understanding the inter-relations of different social and economic characteristics.

3. *How much detail do we sacrifice when we try to identify the general trends?*

We began this analysis with seventeen individual indicators describing selected aspects of the social and economic diversity of the metropolis. More than seven-tenths of the information in these original variables can be distilled into four distinct, independent dimensions – Factors 1 through 4 – that can be understood as composite, compound, or hybrid measures. But how much of the original information *in each single variable* is captured in the four-factor solution? We can answer this question with a quick glance at the "final communality estimates" section of Output 2. For each of the original variables, the communality is the sum of the squared loadings on each of the factors. This is a fancy way of expressing a simple concept: the proportion of the information in the original variable that is captured in the four-factor solution. Notice the high communality estimates for population change (0.81), low-income households (0.85), median household income (0.85), homeownership (0.87), and many of the other variables: even when we distill the seventeen original variables down to only four factors, the resulting synthesis still conveys 87 percent of the original information in the variation of homeownership rates across different neighborhoods. By contrast, our four factors of class and racial-ethnic inequality, family breakup, traditional families, and growth don't capture nearly as much of the variation in the proportion of persons identifying themselves as Black (34.9 percent) or Filipino (50.0 percent). The geographical distributions of these populations are not closely related to any of the other dimensions of neighborhood social composition that we have included in the analysis. Neighborhood and individual *contingencies* apparently play significant roles in the choices and constraints of where to live among the estimated 18,405 people in the metropolitan area who identified themselves as Black, and the 57,025 who described themselves as Filipino.[19]

Exploring the Contemporary Urban Mosaic

> "It is time to explore the effect of these changes upon the contemporary urban social differentiation of western cities in a multivariate, not a single variable context ...."
>
> Wayne Davies and Robert Murdie[20]

We are now in a position to explore, analyze, and interpret more of the contemporary social, economic, and cultural complexity that shapes today's metropolis. If you look carefully at the list of variables used for the 'classical' social area analysis in the previous pages, you will suddenly encounter the realization that explains so much of the history of research in this area. *If we choose to measure neighborhoods in terms of economic, family, and racial-ethnic*

[19] These figures do not come directly from the factor analysis, but from the first column of the raw data worksheet including all of the census variables for the metropolitan area. See http://www.geog.ubc.ca/~ewyly/Private/g350/data/vancma01_1.xls

[20] Davies and Murdie. "Consistency and Differential Impact," p. 46.

characteristics, it should come as no surprise that the results of the factorial ecology support the general proposition that the diversity of the city can be distilled down to a few underlying dimensions of residential structure that broadly correspond to these aspects of society. What you get out of a factorial ecology depends on what you put into it. What you see depends not only on how you look at it, but also on what you choose to measure. The measures you choose will, in a fundamental and inescapable way, create the multi-dimensional world in which the mathematical operations of a factor analysis yield the array of eigenvalues, loadings, and all the rest. This is one of the implications of the common refrain across the humanities and social sciences – that the world we see is in large part a social construction, the result of the ideas and concepts we use to try to perceive it.

The results of an urban factorial ecology depend on the choice of variables to measure the social and spatial diversity of the metropolis: what you get out of the technique depends on what you put in.

There is no single, undisputed basis for deciding what aspects of a complex societal creation like the city should be measured.

We can't measure everything, and unless we have a lot of money and time to obtain the information ourselves, we'll have to rely on data provided by other institutions -- usually, government agencies. Like all methods, then, urban factorial ecology is constrained by the politics of data. The elimination of the long-form Census of Canada for 2011 makes it impossible to update the full, detailed urban factorial ecologies shown here.

But to say that our world (or a city) is socially constructed is the beginning of the conversation, not the end. One reason for the retreat from factorial ecologies after the obsessive wave of research in the 1960s and 1970s can be traced to a crisis of epistemology: since many of those active in factorial ecology were committed to many of the principles of positivism, they recoiled when it became clear that it would never be possible to obtain an objective, 'scientific' theory of urban social structure. The findings of every study depended on the choices, priorities, judgments, and subjective hunches of whomever was involved in setting up the measurement scheme and the statistical analysis. There was no single correct answer. Indeed, when geographers delved deeper into the work of the statisticians who developed and refined the statistical techniques themselves, they realized that it was even difficult to decide on a single "best" method to use when trying to find answers to interesting questions.[21] I can't prove it, but I

[21] This realization was particularly destabilizing in the choice of what "rotation" to use after the decision had been made on how many factors to retain: the software I use provides options for Equamax, Harris-Kaiser case II oblique rotation, Orthomax, Parsimax, oblique Procrustes rotation, Promax, Quartimaz, and Varimax. Don't worry. I don't

suspect that much of the retreat from this style of research came when many positivist geographers realized that a suite of methods they had worked so hard to learn had the effect of undermining the fundamental premises of their epistemology. The role of subjectivity, of course, affects all kinds of statistical analyses; but it is especially pronounced in this area, both because of the nature of the method, and because there is no independent, undisputed basis for deciding what aspects of a complex societal creation like a city should be measured. Over the years, geographers tended to move away from the factorial ecology tradition because it could not provide the correct explanation for complex urban socio-spatial patterns. And fewer people thought it necessary to invest the time required to learn the method.

But the questions we ask are often just as important as the answers we propose. If indeed the world and its cities are socially constructed, then we realize that there is no need to search for a *single* explanation, or to try to determine which model of urban spatial structure is correct. The answers we get will depend on the things we choose to measure. These choices reflect judgment, subjectivity, and creativity. This can be a good thing. In place of epistemological anxiety, we can embrace pluralist methodological and interpretive innovation. The search for a single, correct 'answer' gives way to something much more ambiguous and challenging -- but also more interesting and relevant.

> "The approaches, arguments, and conclusions ... of all factorial ecologies ... cannot be evaluated from the scientistic perspective of positivism, *for their essence is the idea that meaning in any situation has to be learned rather than posited by aprioristic theory*. To understand the how and why of factorial ecology, the perspective of a phenomenological philosophy is required."[22]

This simply means that we have to be very reflective about the meanings and assumptions used in the categories and measures we use to describe the world, because "...reflective knowledge can only be derived dialectically from the interplay of the world of our native experience and the structuring activity of our various perceptual and conceptual orientations"; "Factorial ecology is an ingredient in such a dialectic."[23]

To give you a taste of what is possible, I've undertaken a series of factor analyses that follow the same basic steps as outlined in the "classical" example above -- but with a much broader array of measures of social, economic, and housing-stock characteristics. This broad array of measures -- 66 variables measuring everything from household income, rent, and racial-ethnic diversity to occupational segmentation and the amount of unpaid labor men and women devote to housework, child care, and care for elderly relatives -- draws inspiration from Murdie and Teixeira's chapter, which outlines the complexity of postindustrial social and cultural changes in Canadian urban life. This complexity, though, does introduce certain challenges: the results of

know what most of these are all about, either. Most geographers use Varimax, simply because it's the easiest to understand and interpret. We're geographers, not statisticians. See SAS Institute (1999). *SAS/Stat User's Guide, Version 8.* Volume 1. Proc Factor, pp. 1142-1143. Cary, NC: SAS Institute.

[22] Brian J.L. Berry (1971). "Introduction: The Logic and Limitations of Comparative Factorial Ecology." *Economic Geography* 47(2), Supplement, 209-219, quote from p. 214.

[23] Berry, "Introduction," p. 214, 215.

the factorial ecology can become so rich, so intricate, and so nuanced that it takes a long time to interpret all of the results. Therefore, in addition to the "full" analysis of Vancouver's social mosaic, I've also undertaken analyses that focus on selected themes of urban structure and urban social relations.

The detailed code that defines all of the variables is at
http://www.geog.ubc.ca/~ewyly/g350/nfact06sas.txt

The results of the **Full Vancouver Social Mosaic** are at
http://www.geog.ubc.ca/~ewyly/g350/fact06_all.txt

The results for a subset of variables focusing on **immigration and racial-ethnic diversity** are at
http://www.geog.ubc.ca/~ewyly/g350/fact06_imm.txt

The results for a subset of indicators of **housing construction and development cycles** are at
http://www.geog.ubc.ca/~ewyly/g350/fact06_hsg.txt

The results for a set of variables describing **occupational segmentation and gender roles in family-related work** are at
http://www.geog.ubc.ca/~ewyly/g350/fact06_occ.txt

Maps of the factor scores for tracts across the Vancouver metropolitan area are listed on the relevant section of the course projects page.

Your Job

I would like you to use the results of this analysis to explain and interpret social, economic, and housing conditions in any subset of the many different neighborhoods in the Vancouver metropolitan area. You can choose any subset you'd like to explore; but you should use at least some of the results outlined in this background paper to help guide you in your decision.

You have several options for designing an interesting study.

First, you could undertake a careful comparison of the *overall* factorial ecology in relation to the extensive literature on urban socio-spatial patterns. Look at the summaries in the chapters of Knox and McCarthy and Bunting Filion cited earlier, and perhaps glance at Davies and Murdies' Canada-wide article, to identify key hypotheses and interpretations offered by others who've studied the social fabric of cities in Canada and elsewhere. How do you read the results of the analysis of Vancouver's social fabric today? How do the results of the classical factorial ecology line up with the predictions others have made?

Second, you could focus on a particular question of social transformation or policy debate, and provide an in-depth analysis. If you choose this option, consider using one of the 'focused' analyses described above -- the factorial ecology just with the immigration and racial/ethnic diversity variables, for example, or the housing construction cycle indicators -- to clearly highlight the issues you're most interested in.

You could also focus on a particular *variable*, and examine how it correlates with other neighborhood conditions by scrutinizing how it loads on different factors, how much of its variance can be accounted for by all the other variables (i.e., its communality); you could then look at several of the maps, and search for other kinds of information in books, articles, and newspapers to understand what's going on. On the other hand, you might focus on a particular *factor* if you think it seems to be capturing a significant share of the neighborhood variation in aspects of transport and sustainability: you could examine the policy, organizing, or activist implications of the relation between new housing development and inequalities of income and property ownership.[24]

Third, you could choose a small number of individual neighborhoods, and analyze them in depth. One strategy would be to identify neighborhoods with factor scores at the extremes on whichever dimension you wish to study. Another would be to identify parts of the neighborhood maps where neighborhoods with sharply divergent scores on various factors are situated right next to each other, suggesting a process of dramatic change, a sharp boundary between two different communities, or other features of the social and built environment that create sharp contrasts. Then search out scholarly articles, books, or newspaper sources to explore how the neighborhoods are changing, and how the spatial configuration of this part of the city reflects (or possibly influences) social relations and/or debates over public policy.

There are many other options, and you are encouraged to be creative. You do not need to perform any statistical calculations in your project; on the other hand, I would not recommend that you completely ignore all of the work I've put into this background paper, either. Use the results of the factorial ecology to help you formulate hypotheses, and to choose which socio-spatial issues or neighborhoods to explore.

You may want to consult some of the 'raw' data obtained from Statistics Canada. All of the tract-level measures provided by Statistics Canada for 2006 are split between nine different files, listed on the class project web page. A more manageable subset of the variables I used for the factorial ecology is at:

http://www.geog.ubc.ca/~ewyly/Private/g350/data/vancma06.xls

Data for 2001 are also available, but tract boundary changes and other complications make it difficult to conduct a precise, comparative analysis of neighborhood change.[25]

[24] For a full-length illustration of one way to use the approach to examine a particular aspect of urban socio-spatial patterns, see Ivan Townshend and Ryan Walker's detailed analysis of the dimensions of income segregation in Canadian cities. Ivan J. Townshend and Ryan Walker (2002). "The Structure of Income Residential Segregation in Canadian Metropolitan Areas." *Canadian Journal of Regional Science* 25(1), 25-52.

[25] For 2001, the data are split between two files: http://www.geog.ubc.ca/~ewyly/Private/g350/data/vancma01_1.xls
http://www.geog.ubc.ca/~ewyly/Private/g350/data/vancma01_2.xls
I wish it were possible to include everything in a single file, but one of the frustrating features of Microsoft Excel is that there is a limit to the number of columns that can be included in a single worksheet.

Regardless of which path you choose, you should certainly skim the sources cited above on page 3, if you have not already done so. Once you’ve made a tentative choice on one of the paths outlined above (or any other that seems logical to you), begin to sketch out your notes summarizing which of the aspects of the literature on old and new streams of social area analysis seem most relevant and important for your approach to Vancouver. Then sift through the evidence in the factorial results, using this background paper (as well as the sources cited above on page 3) as a guide to help you interpret things. Also, depending on the path you’ve chosen, you should search for other sources on particular policy issues or neighborhoods, using a judicious mixture of scholarly searches and press searches. Finally, you should draft a paper presenting your findings and interpretations.