Global Analytics
Four Illustrative Stories and a
Virtual Academic Community (VALSAC)

Dominique Haughton, Bentley University and Université Toulouse I

https://sites.google.com/site/dominiquehaughton/home
http://dominiquehaughton.wiki.baihui.com/
dhaughton@bentley.edu

In collaboration with S. Arumugam, Kalasalingam University, India, John Boland, University of South Australia, Irene Hudson, University of Newcastle, Australia, Phong Nguyen, T&C Consulting, Hanoi, Vietnam, Maria Skaletsky, Bentley University, B. Vasanthi, Kalasalingam University, India and Changan Zhang, Bentley University

This paper presents four stories in order to give a sense of what global analytics (AKA data analysis, data science, data mining, applied to international data) entails, and then motivates the VALSAC (Virtual Academic Living Standards Analytics Community) initiative.

The first story takes us to Tamil Nadu (India) and discusses a Social Network Analysis (SNA) of a community of weavers in the village of Sankarapandiapuram.

The second story takes us to Viet Nam. We demonstrate how Kohonen maps - otherwise known as self-organizing maps – can be used to address the issue of comparing the living standards of Vietnamese provinces. In the past single indicators such as the GDP per capita or the poverty rate have been used to rank provinces. However it is recognized that the living standard of a province is a multi-dimensional concept, and the problem arises of how to rank provinces on the basis of several indicators. Attempts to use composite indicators such as the Human Development Index (HDI) carry some limitations. The Kohonen map methodology makes it possible to map Vietnamese provinces on a two-dimensional grid, on the basis of a number of living standards indicators, and to give an interpretation of the dimensions on the grid. Our results shed some light onto some intriguing past rankings based on single or composite indicators, and at the same time demonstrate the usefulness of the Kohonen map methodology in the social sciences.

The third story also takes place in Viet Nam. We discuss a methodology (multilevel models) to obtain small area estimates in the context of the Vietnam Living Standards Surveys. The problem is that household surveys provide excellent and typically unbiased data on household living standards but tend to be expensive to administer and thus have small sample sizes, making estimation at such disaggregated levels as communes so inaccurate as to be impractical. Our models for estimating more accurately the commune-level mean (logarithm of) household expenditure per capita rely on independent variables available both in the 1999 Census and in the VHLSS of 2002. As a useful by-product, we also obtain a *Location Impact Factor (LIF)* for each commune, a measure of the impact on living standards of commune location when characteristics such as the education and employment structure etc of the commune population are controlled for.

The fourth story is set in Russia. It leverages the multilevel models methodology to generate the first comprehensive investigation of the digital divide in Russia, on the basis of recent household survey data.

Inspired by these four stories, the paper then motivates and describes the VALSAC (Virtual Academic Living Standards Analytics Community) initiative.

Story 1: Reciprocity in social networks - A case study In Tamil Nadu, India (with S. Arumugam, B. Vasanthi and Changan Zhang)

Introduction

Reciprocity refers to responding to a positive action with another positive action; it creates, maintains and strengthens various social bounds. It is the foundation of social order and is a major key to success. This applies not only in social networking but also in all rounds of human activities. The potential for reciprocal actions by players increases the rate of contribution to the public good; reciprocity is a form of social obligation and is a motivation for returning favors from others (Fehr et al. 2000). Reciprocity was studied and evaluated from the beginning of social network analysis in the 1930's. A measure of reciprocity is a number which gives the extent to which support is both given and received in a relationship.

Reciprocity and social capital

The investigation of social networks such as the ones in this story is important from the *social capital* point of view. As stated by Claridge (www.socialcapitalresearch.com), "social capital is about the value of social networks, bonding similar people and bridging between diverse people, with norms of reciprocity" (Dekker and Uslaner 2001; Uslaner 2001). Social capital in turns is of importance to economic development, an idea which has spawned a considerable literature, dating in large part from the early 2000s.

Network data

The population of our social network study is a small closed set of actors consisting of 100 well organized households in the small village of Sankarapandiapuram in Tamil Nadu, India. This village has just four streets named North Street, South Street, Kallakudi Street and Pallakudi Street (see Figures 1a, 1b and 1c). All the members of the various households under consideration belong to the same community called "Saliyar", which is considered to be a poor community in the state of Tamil Nadu. The basic business of this community is weaving. During the past two or three decades, several members from this community have opted for higher studies and are employed in several posts such as engineers, doctors, teachers, but more than 80% of this community are engaged in weaving, either with a hand or power loom and depend on their daily earning for their livelihood. Most members of the community would be considered to lie below the middle class category in India.

Most of the respondents in this study work in surgical cotton industry, the main manufacturing product being bandage clothes, which are exported to several countries. All the households under consideration are closely located and interact among themselves almost on day-to-day basis. We have collected data from a hundred households through a questionnaire and personal interview. The network data include the name and age of the head of the household and his wife, the educational qualifications of the head, the number of dependents in the family and their employment details. The 100 households are labeled with the numbers 1, 2, …, 100; for each household i we have data consisting of the list of households whom they approach for monetary help, advice and companionship for spending leisure time, both during crisis and normal periods. The data yield six directed graphs on the set of nodes $\{1,2,\ldots,100\}$. Apart from the above data we also know the list of relatives and (mutual) friends for each household i, which give two undirected graphs on the same vertex set.

Figure 1a: General location of the village of Sankarapandiapuram in India

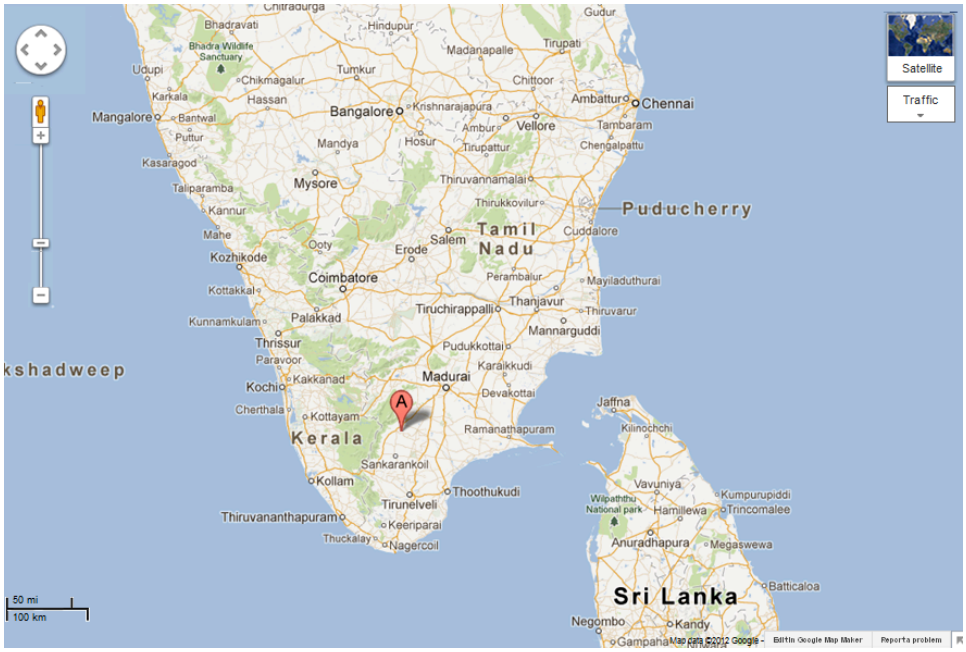


Figure 1b: Map of the village of Sankarapandiapuram in India

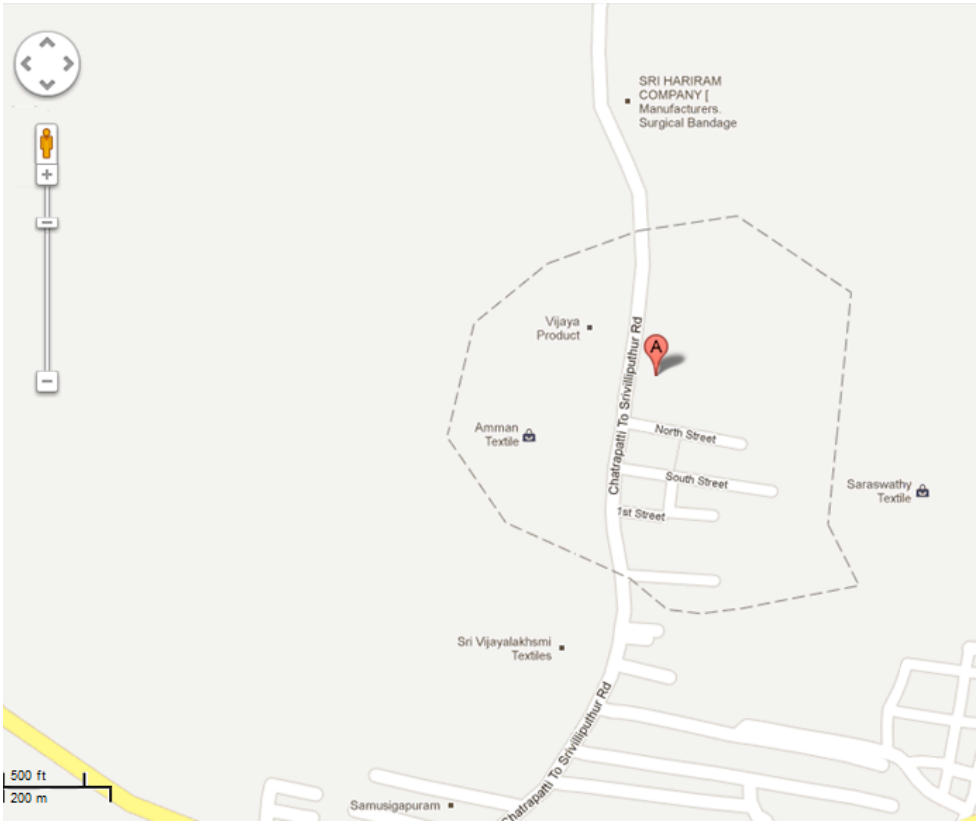


Figure 1c: Satellite view of the village of Sankarapandiapuram in India

Let D_1 (D_2) be the directed graph representing the network of monetary help during crisis (normal) periods. Let D_3 (D_4) be the directed graph representing the network of advisory help during crisis (normal) periods. Let D_5 (D_6) be the directed graph representing the network of companionship during crisis (normal) periods. Figures 2a-b, 3a-b, 4a-b display the 6 networks.

Reciprocity in the network

In networks D_1 and D_2 representing monetary help during crisis and normal periods, there are six and four reciprocal ties respectively; it is interesting to note that all these reciprocal ties are within relatives. In networks D_3 and D_4 representing advisory help during crisis and normal periods, there are 12 and 12 reciprocal ties respectively and in both cases 10 of the reciprocal ties are within relatives. However, the reciprocity behavior is different in networks D_5 and D_6 representing companionship. In network D_5 there are 38 reciprocal pairs and out of these, 21 are between relatives and 17 are between friends. In network D_6 there are 46 reciprocal ties and out of these 25 are between relatives and 21 are between friends. Thus respondents have mutual reciprocal interaction outside their circle of relatives only for companionship during leisure time. Table 1 lists the reciprocity measure for each network, equal to the proportion of links which are bi-directional.

Table 1. Reciprocity measures for each network.

Monetary		Advice		Companionship	
Crisis	Normal	Crisis	Normal	Crisis	Normal
D_1	D_2	D_3	D_4	D_5	D_6
.13	.10	.15	.14	.27	.33

It is clear that reciprocity is quite a bit higher in the companionship network. The difference in reciprocity in crisis and normal times is modest in general, except possibly for the companionship network, where normal times seem to encourage reciprocity.

Figure 2a. Monetary help in crisis periods (network D_1)

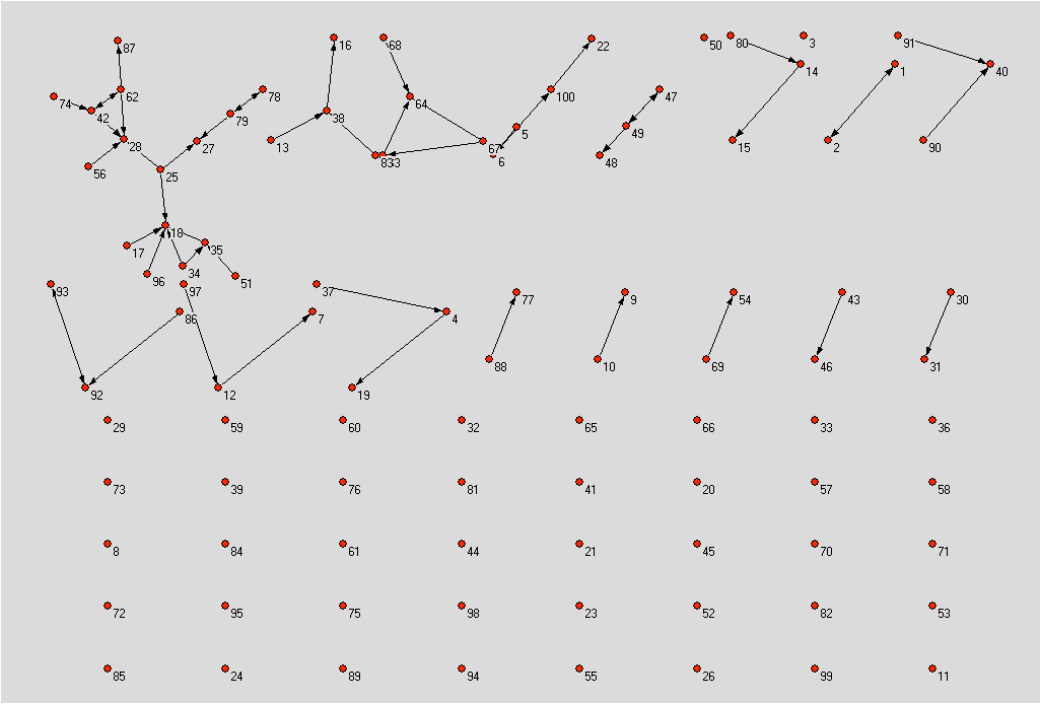
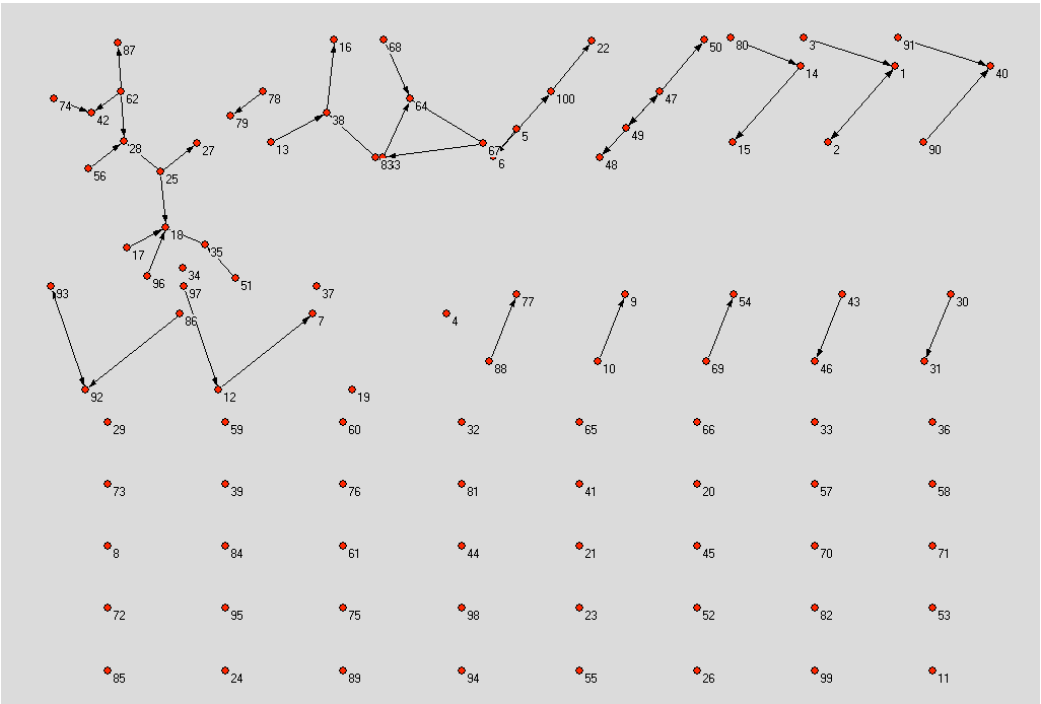


Figure 2b. Monetary help in normal periods (network D_2)



In the monetary help networks, we observe that (42, 62) and (78, 79) are reciprocal pairs during crisis periods, but are not reciprocal during normal periods. In fact, 42 approaches 62 for monetary help only during crisis periods whereas 62 approaches 42 for monetary help both during crisis and normal periods. The same situation prevails for the pair 78, 79; here 79 approaches 78 for help only during crisis periods. There is another interesting similarity between the pairs (42, 62) and (78, 79). The respondents corresponding to each of these pairs are close relatives (father/son relationship in one case and brother/sister relationship in the other case).

In-degree and out-degree

In a directed network, the *in-degree (id)* of a vertex is defined to be the number of arrows directed to the vertex and the *out-degree (od)* of a vertex is the number of arrows which arise from the vertex. The maximum in-degrees in D_1 and D_2 are respectively 5 and 4; respondent number 18 has maximum in-degree in both D_1 and D_2. He is the owner of an industrial plant and is active in politics. He is also the village head and is naturally the most influential person in networks D_1 and D_2.

Out of the 100 respondents, 65 have in-degree 0 in D_1 and 68 have in-degree 0 in D_2. This is perhaps not surprising since most of the respondents under consideration lie just above the poverty line and hence are not in a position to provide monetary help to others, so that no one approaches them for monetary help. Also the maximum out-degree of a vertex both in D_1 and in D_2 is 3. Respondents 25 and 62 have out-degree 3 in D_2. This shows that exchange of monetary help is very minimal in the network (see Figures 2a and 2b). On the other hand, 59 respondents have out-degree 0 in D_1 and 63 respondents have out-degree 0 in D_2; this shows that a large proportion of the respondents seem to be able to cope with the limited income they earn. Perhaps this is typical of any small Indian village.

Respondent number 1 has maximum in-degree in D_3 and D_4; he is educated and is a manager in a textile export company; his wife is a tailor who produces garments intended for ladies and is an active member of the women's self-help group in the village. Respondent number 11 has maximum in-degree in D_5 and D_6; he is an astrologer.

Let D=(V,A) be a directed graph. A vertex $v \in V$ is called

(i) an isolated vertex if od(v) = id(v) =0
(ii) a transmitter if od(v)>0 and id(v)=0
(iii) a receiver if od(v)=0 and id(v)>0
(iv) a carrier if od(v)>0 and id(v)>0

The distribution of the 100 vertices in various categories is given in Table 2.

Table 2. Distribution of the respondents across the various categories

Network	Isolated	Receiver	Transmitter	Carrier	Max out-degree	Max in-degree
D_1	42	17	23	18	3	5
D_2	44	19	23	14	3	4
D_3	21	21	25	33	3	8
D_4	19	21	25	35	4	7
D_5	4	4	14	78	6	8
D_6	4	4	13	79	6	8

Figure 3a. Advisory help in crisis periods (network D_3)

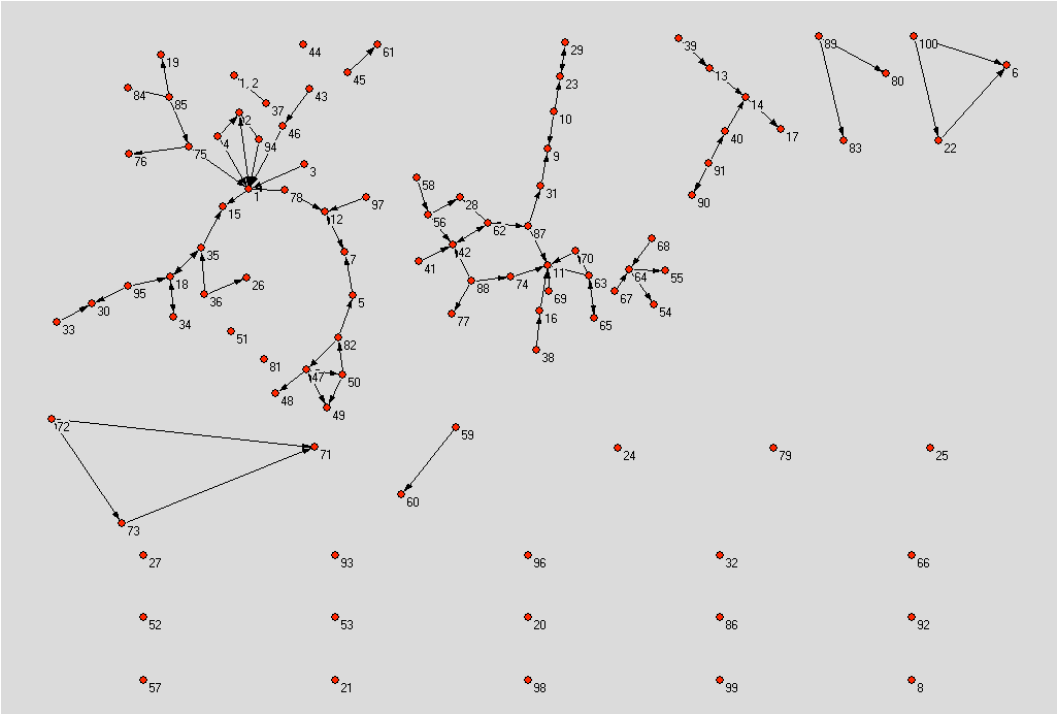
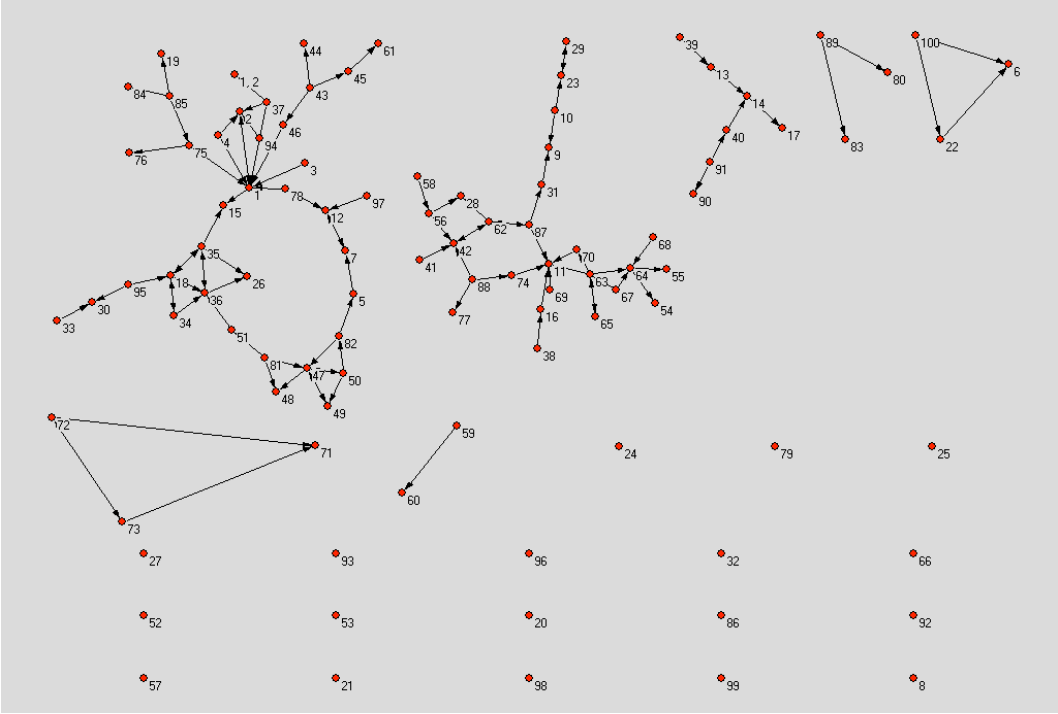


Figure 3b. Advisory help in normal periods (network D_4)



Note that no respondent is isolated in all six networks. Respondent number 8 has in-degree 0 in all six networks and has out-degree 0 in all networks except in D_5 and D_6; and in these networks the out-

degree is 3. All three out-neighbors of this respondent in D_5 and D_6 are his relatives. Thus no respondent approaches 8 for any type of help.

Figure 4a. Companionship in leisure time in crisis periods (network D_5)

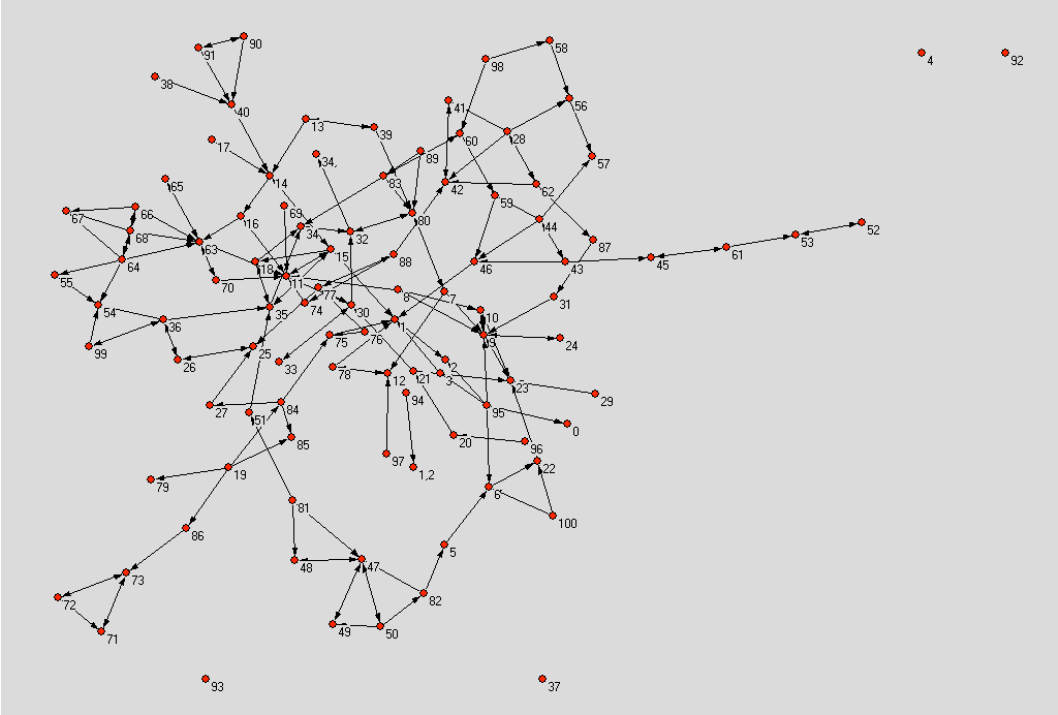
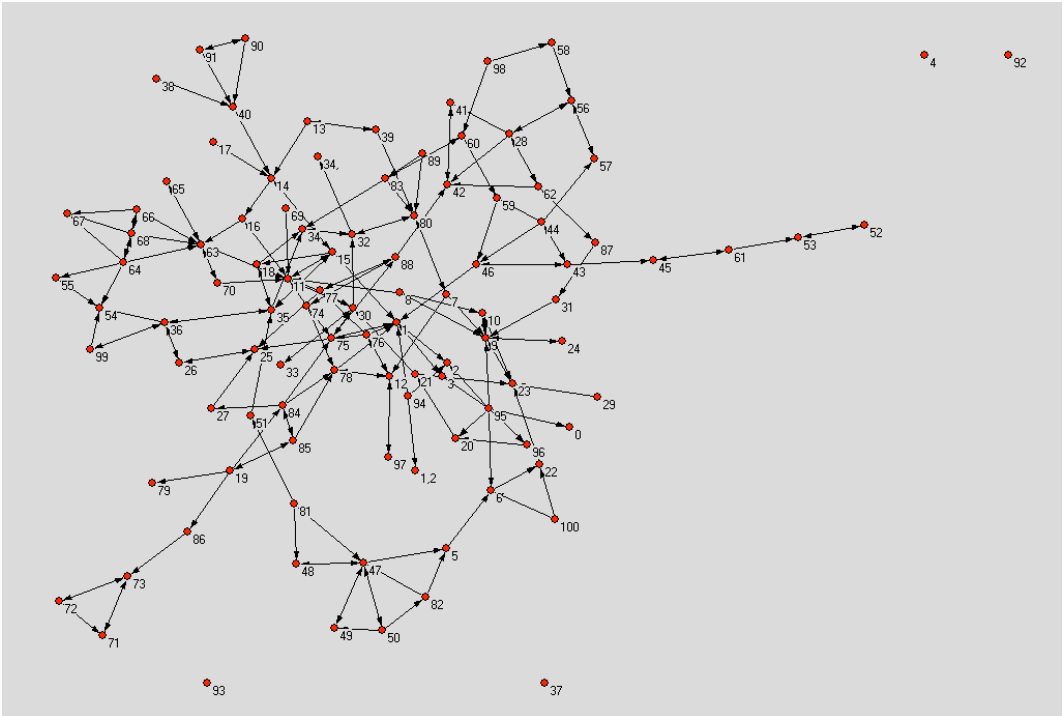


Figure 4b. Companionship in leisure time in normal periods (network D_6)



When we compare isolated vertices in networks D_1 and D_2 representing monetary help, we observe that vertices 3 and 50 are isolated in the crisis network but not isolated in the normal network. Also respondents 4, 19, 34 and 37 are isolated in the normal monetary help network and are not isolated in the crisis network. Thus these respondents seek monetary help only during crisis and otherwise they are able to manage on their own.

Connected Components

The number of nontrivial components in network D_1 representing monetary help during crisis periods is 16. The number of vertices in the largest component is 16; vertices 18 and 28 play an important role in providing financial help for members of these components. There are six components each with 2 vertices. The number of vertices in the largest component in network D_2 is 13; here also 18 and 28 have significant contributions. In the network of advisory help there are two large components with 31 and 22 vertices in D_3 and 32 and 27 vertices in D_4. The other non-trivial components are relatively smaller. In the networks of leisure time companionship during crisis and normal period, there is a single giant component which contains 96 and 93 vertices, respectively, which indicates that the members of the community as a whole have reasonably good interaction with each other.

In the companionship network, respondents 4, 37, 92 and 93 are isolated. We observe that for these respondents, isolation is a matter of personal choice. For example respondent 37 is an old woman living alone with monetary help from her sons who has no inclination for mingling with others. Similarly for other personal reasons the remaining three respondents have chosen to isolate themselves from the rest of society and do not entertain visitors.

To conclude

This story has painted a picture of a community of weavers in a small Tamil Nadu village from the lens of social network analysis (SNA) and has identified subgroups and influential actors in the community. Several interesting questions arise from this study, for instance: which type of social structure might tend to lead to higher living standards for the community? Do linkages tend to differ significantly in crisis and normal times? Both these questions give rise to interesting and challenging statistical problems.

References: Story 1

Dekker, Paul, and Eric M. Uslaner. 2001. 'Introduction.' Pp. 1 - 8 in *Social Capital and Participation in Everyday Life,* edited by Eric M. Uslaner. London: Routledge.

Fehr, Ernst and Simon Gadster (2000), Fairness and Relation; The Economics of Reciprocity, Journal of Economic Perspectives, 14(3), 159-181.

Uslaner, Eric M. 2001. "Volunteering and social capital: how trust and religion shape civic participation in the United States." Pp. 104 - 117 in *Social Capital and Participation in Everyday Life*, edited by Eric M. Uslaner. London: Routledge.

Story 2: Living standards of Vietnamese provinces: a Kohonen Map (with Phong Nguyen and Irene Hudson, excerpt from CSBIGS 2(2))

The measurement of living standards is widely recognized to be a multivariate challenge. In this story, problems with the use of typical univariate indicators are outlined, and a novel approach is suggested which relies on Kohonen maps. The stability and accuracy of the map are evaluated via a bootstrap methodology. This story presents an application of the technique of Kohonen maps in the context of a data set of indicators about Vietnamese provinces.

Introduction

In Viet Nam where provinces have been competing with each other in the area of economic development and fast poverty reduction, government leaders, policy makers and managers as well as researchers usually ask the question: "Which province is better off?" Efforts in answering this question lead to another question: "Which indicator should be used to rank provinces?". Traditionally, one used single indicators such as the Gross Domestic Product (GDP), the household income per capita, or the poverty rate in order to rank provinces. The report "National Human Development Report 2001: Doi Moi and Human Development in Viet Nam" (NHDR 2001) issued under the leadership and coordination of the National Center for Social Sciences and Humanities with the support of UNDP (United Nations Development Program) (National Political Publishing House, 2001) proposed a ranking of the 61 provinces of Viet Nam from most developed (1st position) to least developed (61st position) on the basis of the HDI – the Human Development Index.

The four indicators we have just mentioned, both single and composite, give different rankings of provinces. As an illustration, focusing on the GDP per capita in 1999, the household income per capita in 2002, the poverty rate in 2002, and the HDI in 1999, the rankings of four specific provinces are displayed in Table 1 below. Ha Noi is the capital city of the country and one of its most developed cities. Ho Chi Minh city is the largest and the most developed city in Viet Nam. Ba Ria Vung Tau is a province with natural oil and attractive beaches popular with tourists. Binh Duong is a new industrial province.

Table 1. Rankings of provinces according to different indicators

	GDP per capita 1999	Household Income per Capita 2002	Poverty Rate 2002	HDI 1999
Ha Noi	3	2	5	2
Ho Chi Minh City	2	1	1	3
Ba Ria Vung Tau	1	3	7	1
Binh Duong	4	6	2	6

In this case the question arises of which indicator should be used. In Viet Nam the HDI seems to be the preferred indicator in ranking provinces. Experts in government circles argue that because living standards are multidimensional, a composite indicator such as the HDI should be more suitable than any single indicator.

The HDI was developed and used by UNDP in ranking countries in terms of levels of human development. The HDI measures average achievements in a country in three basic dimensions of human development: (i) a long and healthy life, as measured by life expectancy at birth; (ii) knowledge, as measured by the adult literacy rate (with two-thirds weight) and the combined primary, secondary and tertiary gross enrolment ratio (with one-third weight); and (iii) a decent standard of living, as measured by the GDP per capita (in PPP - Purchasing Power Parity - US $).

HDI rankings in the report were felt in Viet Nam to be reasonable for Ha Noi (2nd position), HCM city (3rd position), and Binh Duong (6th position), but the first position allocated to Ba Ria Vung Tau by the HDI was considered by many to be very counter-intuitive. It was felt that the high GDP per-capita of Ba Ria Vung Tau, mostly from natural oil, dominated the HDI value for the province. As a result, this province was not mentioned in the analysis part of the report, leading among other things to dissatisfaction in several quarters.

We propose here to use the technique of Kohonen maps (Kohonen, 2001) as an alternative to this state of affairs. Kohonen maps (also known as Self Organizing Maps) have been used in the area of living standards and poverty analysis. Albert et al. (2003) used a Kohonen map with 15 poverty indicators to identify poor provinces in the Philippines for government poverty intervention. Kaski and Kohonen (1996) used 39 welfare indicators and a Kohonen map to compare the economic level and the standard of living of different countries. Ponthieux and Cottrell (2001) used the Kohonen algorithm to combine different measures of living conditions and to classify households by their level of living conditions as well as differences within similar levels. We also refer the reader to Deichman et al. (2007) for a study of the international digital divide which relies on Kohonen maps and includes a fairly extensive discussion of the Kohonen methodology.

In this past work, however, no method is proposed or applied for validating the reliability of the Kohonen map. We will use a bootstrap approach and statistical tools to assess the reliability of self-organizing maps, following ideas in De Bodt et al (2002).

Kohonen Map Methodology: a brief introduction

Kohonen maps, due to T. Kohonen and his research team in Finland, are a special case of a competitive neural network, and are also referred to as Self Organizing Maps (SOMs). The web site http://www.cis.hut.fi/projects/somtoolbox/ contains a useful introduction to the methodology and a Matlab 6.0 toolbox (which was used here) to build the maps.

The basic algorithm for constructing Kohonen maps is as follows:

i. Begin with a grid, typically 2-dimensional, with a vector $m_i(t)$ assigned to each grid position, initially typically randomly, of the same dimension as the number of variables.
ii. For each data vector x(t) find the best match c on the grid such as:

For every i, $| x(t) - m_c(t) | \leq | x(t) - m_i(t) |$

iii. Update the vectors $m_i(t)$ as follows:

$$m_i(t+1) = m_i(t) + h_{c,i}(t).(x(t) - m_i(t)),$$

where $h_{ij}(t)$ is the neighborhood function, a function of t and of the geometric distance on the lattice between position *i* and position *j*. Typically $h_{ij} \to 0$ as the distance between *i* and *j* increases and as more iterations are performed.

iv. Iterate this step over all available data vectors, and repeat until little change is observed in the $m_i(t)$.

The resulting map tends to organize the components of the estimated vectors, the $m_i(t)$, in a monotonic way (increasing or decreasing) as one moves on the map, hence the term *Self Organizing Maps*.

Kohonen Maps and Vietnamese provinces

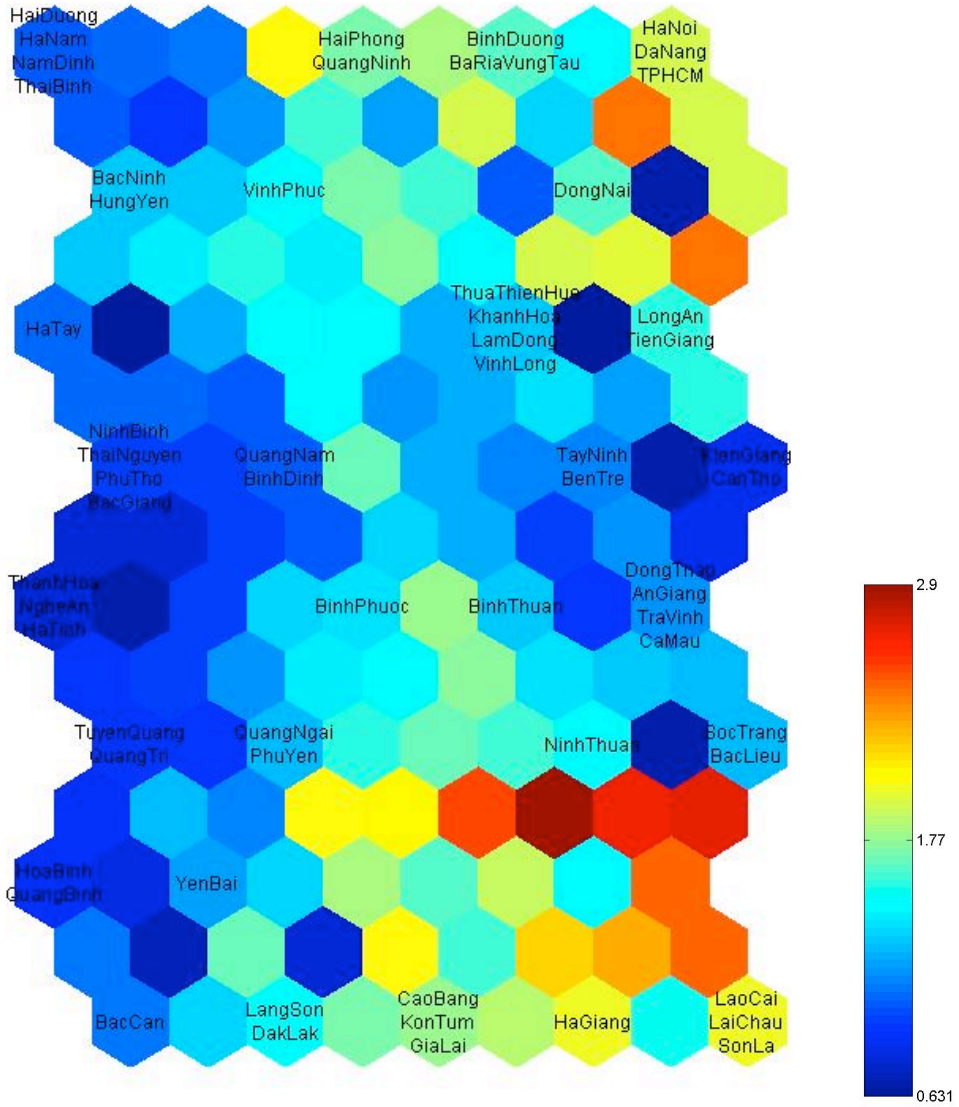
We decided to use the set of 25 variables listed in Table 2 in order to represent with the same number of variables (eight) each area covered by the HDI index, wealth, education and health, with an additional variable on household size.

Table 2. Indicators used in the Kohonen map

Wealth of household	Education level of household	Health of household
Gdpindex99: GDP Index normalized between 0 and 1, and truncated (ppp 1999 values), UN index *Hpiindex99:*UN composite poverty index (1999) *Gdppc99vnd:* GDP per cap. in 99 VND *Mincpccurpric:* Avge income per cap. (current prices) *Percexpfood:* % exp. allocated to food *Valdurgoods:* Avge value of durable goods *Percpermhouse:* % permanent houses *Povrate:* Poverty rate	*Eduindex99:* literacy rate (2/3) and combined enrollments rates (1/3), normalized between 0 and 1, UN composite index *Adlircy*: Adult literacy rate *Percvocswc*: % with vocational sec. school diplomas *Percunivcoll*: % with university and college diplomas *Percmsphd*: % with Masters and PhD degrees *Schlen3to5*: School enrollment rate ages 3-5 *Schlenlowsec*: Lower secondary school enrollment rate *Schlenupse:* Upper secondary school enrollment rate	*Lifexpindex99*: Life expectancy normalized between 0 and 1 (1999 values), UN index *Malnuunder598*: Under five malnutrition rate *Lifexpmale*: Life expectancy (male) *Lifexpfem*: Life expectancy (female) *Wgtforht*: Weight for height malnutrition *Htforage*: Height for age malnutrition *Wgtforage*: Weight for age malnutrition *Matdeathp1000*: Number of pregnancy related deaths per 1000 *Infmort*: Infant mortality rate
Hhsize: Number of members of household		

Source: NHDR 2001 (National Political Publishing House 2001), and Figures on Social Development, Doi Moi in Vietnam (Statistical Publishing House, 2000).

Figure 1. U-matrix of the Kohonen map of Vietnamese provinces on the basis of 25 indicators



The map is two-dimensional (8 rows by 5 columns, as chosen by default by the Kohonen Matlab toolkit software) and each of the 40 positions is associated with a 25-dimensional estimated component vector, obtained at convergence of the Kohonen algorithm. Each of the 40 positions is represented by a hexagon on a display referred to as the U-matrix, with additional hexagons added around each actual position hexagon. Hexagons surrounding an actual map position display different colors to represent the distance to other map positions; the color of a position hexagon corresponds to the average distance between this hexagon and its neighbors. For instance, the top right hexagon (HaNoi, DaNang and TPHCM) has an estimated vector which is not too different from that of the hexagon with BinhDuong and BaRiaVungTau (pale blue) and is moderately different from that of its neighbors (pale green). When the map is built, provinces are placed on it according to the estimated vector their data vector is closest to.

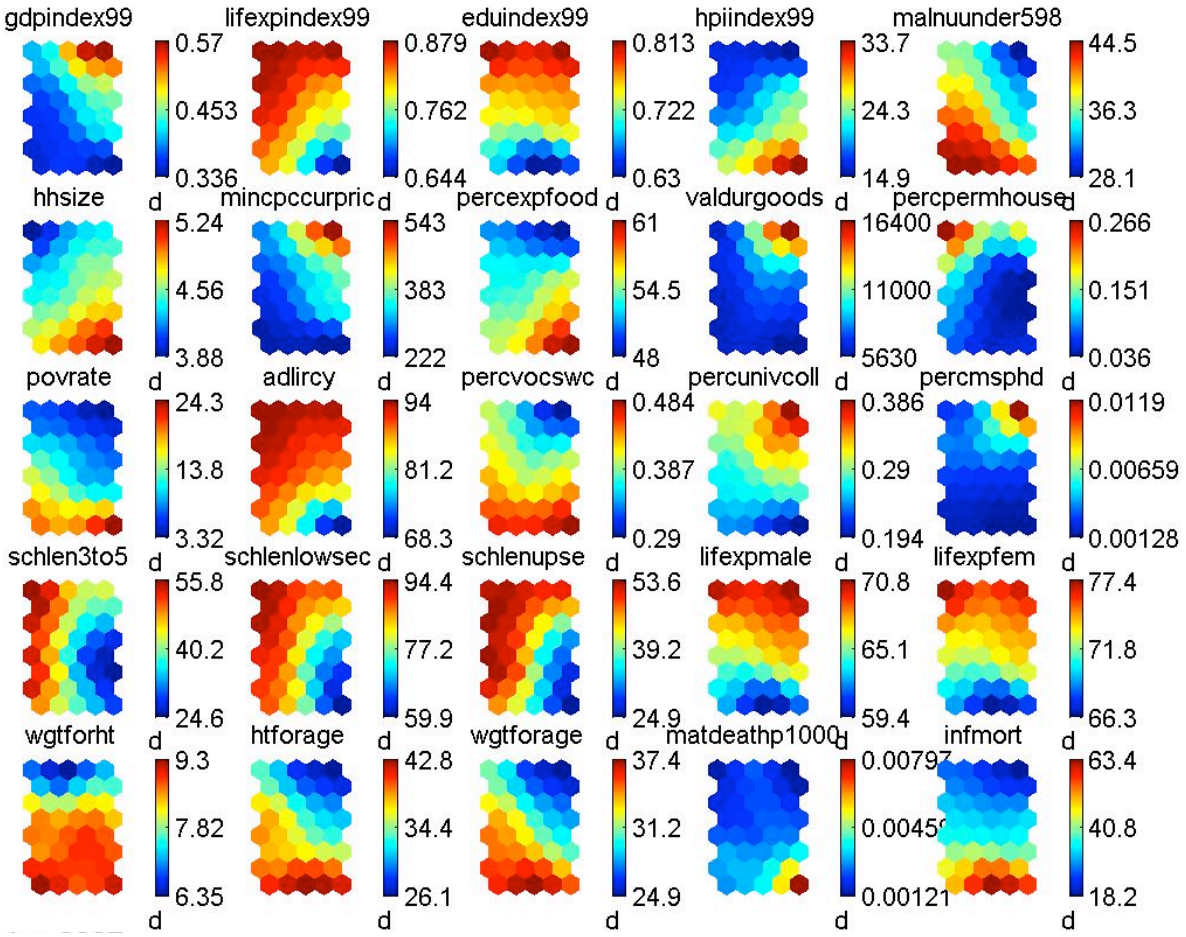
We can see from the map that Vung Tau, which is ranked first by HDI, now is clustered by the two-dimensional Kohonen Map with Binh Duong, which is ranked 6th by HDI; this is a more credible result if

one takes into account area knowledge about these provinces.

We tried a one-dimensional Kohonen map with the same 25 variables. This map shows that Ba Ria Vung Tau has the third position.

Figure 2 displays the estimated values of each of the 25 components used to build the map. The self-organizing property of the map is clearly visible, and the diagonal directions seem to essentially represent a wealth and health axis, and an education axis. For example, estimated GDP per capita can be seen to decrease from the top-right to the bottom left part of the map, while school achievement variables tend to decrease from the top-left to the bottom-right part of the map. One feature to note is that the position of each of the 40 "map position" hexagons is the same on each component map and on the U-matrix.

Figure 2. Components of the Kohonen map of Vietnamese provinces (25 indicators)



We will use a bootstrap approach and statistical tools to assess the reliability of self-organizing maps, following ideas proposed by De Bodt et al (2002).

We created 111 samples of size 61, by random selection with replacement from our initial sample of 61 provinces, and with the stipulation that Ba Ria Vung Tau and Binh Duong (the two provinces we

will focus on for studying the neighborhood stability of the map) should be part of the samples. Each of the 111 samples is referred to as a bootstrap sample.

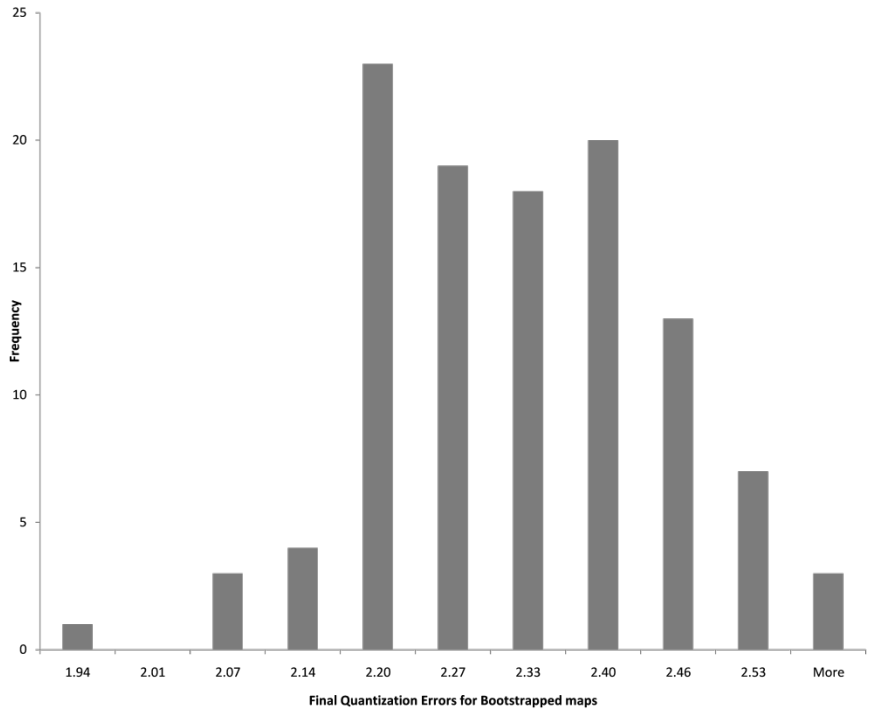
We first assess the stability of the quantization in the Kohonen map (which was estimated on the basis of standardized data) using

$$\text{CV(SSIntra)} = 100\frac{\sigma_{SSIntra}}{\mu_{SSIntra}}$$

where SSIntra denotes the sum of squares of quantization errors – that is the squared distance between an observed data vector x^i and its corresponding (nearest) unit at convergence on the Kohonen map – computed over all observations in each sample, $\mu_{SSIntra}$ denotes the mean of SSIntra errors of all bootstrapped samples, and $\sigma_{SSIntra}$ denotes the standard deviation and the SSIntra errors. A small value of CV implies that the SSIntra values are stable.

For our original Kohonen map, SSIntra (also often referred to as quantization error, as for example in Matlab) is 2.481. For bootstrapped maps $\mu_{SSIntra}$ = 2.290, $\sigma_{SSIntra}$ =0.125 and CV = 0.055. This implies a good stability of the value of SSIntra. A histogram of the bootstrapped values of SSIntra is displayed in Figure 3.

Figure 3. Histogram of quantization errors (SSIntra values) for bootstrapped maps



Second, we assess the stability of the neighborhood relations in the Kohonen map. To do so we assess the stability and its significance of the neighborhood for Binh Duong and Ba Ria Vung Tau as a specific pair of observations.

1 if x^i and x^j are neighbors within radius r

For assessing the stability of the neighborhood for Binh Duong and Ba Ria Vung Tau, we will calculate:

$$STAB_{ij}(r) = \frac{\sum_{b=1}^{B} NEIGH_{ij}^{b}(r)}{B}$$

where $NEIGH_{ij}^{b}(r)$ =

Being neighbors within radius *r* means that the two observations are projected on two centroids on the map such that the distance between these centroids is smaller than or equal to *r*. If *r* equals 0, the two observations are projected on the same centroid; if *r* = 1, the two observations are projected on the same centroid or on immediately neighboring centroids.

Note that the superscript *b* denotes bootstrap sample *b*.

For the pair of Ba Ria Vung Tau and Binh Duong and for *r* = 0, $STAB_{BaRiaVungTauBinhDuong}(0)$ = 0.4324. This implies that 43% of bootstrap samples have Ba Ria Vung Tau and Binh Duong projected on the same centroid. For *r* = 1, $STAB_{BaRiaVungTauBinhDuong}(1)$ = 0.8829. This implies that 88% of bootstrap samples have Ba Ria Vung Tau and Binh Duong projected on the same centroid or on immediately neighboring centroids.

The significance of the neighborhood for a pair can be evaluated by the standard deviation of the proportions .4324 and .8829 (respectively .0470 and .0305).

Conclusion

This story used a study of living standards in 61 Vietnamese provinces to demonstrate that the Kohonen map methodology can serve as a better tool to rank provinces, compared to measures such as the Human Development Index (HDI). It makes it possible to take into account a wide range of indicators in the ranking, and to give more sensible results. In general, the technique is also useful for mapping for example geographical areas like provinces into similar clusters onto a two (or one)-dimensional map on the basis of a larger number of socio-economic variables.

References: Story 2

Albert, J.R., Elloso, L., Suan, E. & Magtulis, M.A., 2003. Visualizing Regional and Provincial Poverty Structures via the Self-Organizing Map. *The Philippine Statistician*, 52 (1-4), p. 39-57.

De Bodt, E., Cottrell, M. & Verleysen, M., 2002. Statistical Tools to Assess the Reliability of Self-organizing Maps. *Neural Networks*, 15, p. 967-978.

Deichman, J., Eshghi, A., Haughton, D. Sayek, S. and Woolford, S., 2007. Measuring the international digital divide: an application of Kohonen self-organising maps*. International Journal of Knowledge and Learning*, 3(6), p. 552-575.

Kaski, S. & Kohonen, T., 1996. Exploratory Data Analysis by the Self-Organizing Map:

Structures of Welfare and Poverty in the World. *Neural Networks in Financial Engineering*, p.498-507. Singapore: World Scientific.

Kohonen, T., 2001. *Self-Organizing Maps,* 3rd edition. Berlin: Springer-Verlag

National Political Publishing House, 2001. Doi Moi and Human Development in Vietnam. Available at http://hdr.undp.org/docs/reports/national/VIE_Vietnam/Vietnam_en.pdf.

Ponthieux, S. & M. Cottrell., 2001. Living Conditions: Classification of Households Using the Kohonen Algorithm. *European Journal of Economic and Social Systems*, 15(2), p. 69-84.

Statistical Publishing House, 2000. Figures on Social Development: Doi Moi Period in Vietnam.

Story 3: Living standards and Location Impact Factors (LIF) in Vietnam (with Phong Nguyen, John Boland and Irene Hudson, excerpt from The Future of Business Education)

This story also takes place in Vietnam and is concerned with living standards, using the typical economic standard of annual household expenditure per capita. It is a story which lives in the realm of data analysis, geography, economics and global studies, cutting across the boundaries between topics which might be considered as business topics (such as data analysis and economics) and topics which might be considered as Arts and Sciences subjects (such as geography and global studies). The study of living standards is related to explorations of business issues, since buying power is necessary for business activity to take place.

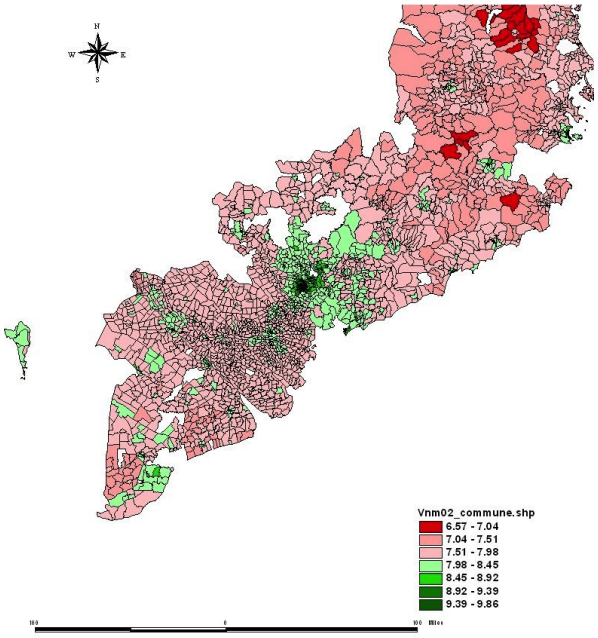


Figure 1a. Small area estimates of living standards in communes in the South of Viet Nam

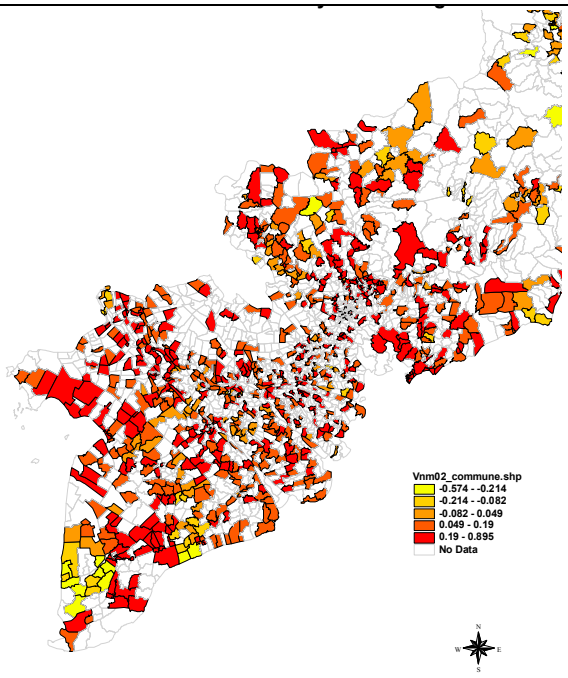


Figure 1b. Location Impact Factors (LIF) for communes in the South of Viet Nam

Note: In Figure 1a, communes in the Mekong Delta seem to display low living standards, but the Location Impact Factors for these communes tend to be high (Figure 1b).

One challenge in this situation is that surveys that seek to evaluate household expenditures are expensive to run; as a result, sample sizes tend to be rather small, and in particular too small to estimate average living standards at the commune level with a decent level of precision. Techniques that combine administrative data (such as census data) with data from the survey can be pressed into service to obtain what is referred to as small area estimates of average living standards at the commune level. The result of this estimation is displayed for southern Vietnam on Figure 1a, where it is clear that communes in the Mekong delta region (in the lower left part of the map) tend to display modest living standards.

However, an interesting insight emerges from Figure 1b, which now displays 'Location Impact Factors' (LIFs) for communes in southern Vietnam that were covered by the living standards survey. These LIFs (Nguyen et al 2012) represent the effect of the geographic location of the commune, when

characteristics such as education, labor structure, household size, and other population factors are controlled for. The LIF for most of the communes in the Mekong Delta on Figure 1b are at the high positive end of the spectrum, implying that the mere location of these communes is associated with a living standards upwards shift.

What is going on here is that the usual correlates of wealth, such as small household sizes, or a high level of education, are not favorable in the Mekong Delta. To some extent, they are moderated by the geographical location of the communes, probably by the presence of waterways and a reasonably favorable climate. Unraveling this insight involves some advanced modeling (Nguyen et al 2012, and Haughton and Haughton 2011), but fundamentally relies on looking at the data with a particular twist.

References: Story 3

Haughton D. and J. Haughton (2011) Living Standard Analytics, Springer.

Nguyen, P., D. Haughton, I. Hudson and J. Boland, Multilevel Models and Small Area Estimation in the Context of Vietnam Living Standards Surveys, Preprint 2012, under revision for *Journal of Data Science*.

Story 4: Application of multilevel models to the study of the individual digital divide in Russia (with Maria Skaletsky)

Introduction

The Digital Divide has emerged as an important research and policy issue during the past thirty years. The digital divide can be defined as an inequality in access to Information and Communication Technology (ICT), such as personal computers, Internet and mobile phones (Norris, 2001). The consequences of the divide are inequality in the ability to obtain important information, which in turn leads to inequality in the ability to gain employment, participate in online communities, e-government, receive important health information and so on. The digital divide is characterized by the disadvantage in the use of and access to ICTs by racial minorities, women, individuals in older age groups and people with low income and low levels of education.

While extensive research already exists on this subject, most existing quantitative digital divide studies are limited to descriptive statistics or simple linear models. Vehovar et al (2006) discuss the lack of sophisticated statistical analysis in digital divide research. Most digital divide studies simply compare the use of ICTs across different demographic groups and are very descriptive in nature. The authors argue that the use of multivariate models and other more advanced techniques would provide a better understanding of the digital divide phenomenon.

We provide a detailed analysis of the state of the individual and regional digital divide in Russia. This is the first comprehensive study that addresses the problem of the digital divide in Russia. We use longitudinal data from the Russian Longitudinal Monitoring Survey (RLMS), a comprehensive living standards survey covering every year from 1992 to 2010, and we rely on the most recent data available for the year 2010. We employ multilevel models to analyze the digital divide problem in Russia at the individual level. Multilevel models allow us to capture geographical effects and account for the hierarchical structure of the data.

Prior research – Individual digital divide

Research on the digital divide in Russia is mostly qualitative in nature or includes basic descriptive statistics only. The same predictors that were found significant in studies of the digital divide in the USA, such as gender, age, and income, are identified as predictors of the digital divide in Russia (Acilar, Markin, & Nazarbaeva; Beketov, 2009; Delitsin, 2006; Lihobabin, 2006). Individuals with lower income, older people and women have lower access to ICTs. In addition to these common problems, a language barrier affects the ability to use Internet content for most Russians – most Internet content exists in English and is inaccessible to most Russians (Beketov, 2009). Another finding specific to the situation in Russia is that there is a very significant regional divide (Beketov, 2009; Delitsin, 2006). The difference in the rate of use of ICTs is between regional capitals and the areas outside the capitals, with the highest rates of the ICT use concentrating in Moscow and Saint-Petersburg.

We build an explanatory multilevel model of the digital development at the individual level. This model provides detailed information about the effects of each of the indicators included in the study, accounts for the hierarchical nature of the data and includes regional effects. We are also able to identify regional effects for indicators included in the model, e. g. we are able to determine whether or not the gender effect differs across regions.

Preliminary results

As expected, we find age, gender, education, income and ability to speak a foreign language to be significant predictors of PC use. Holding all other parameters constant, age negatively affects PC use, while being a male increases it. Higher income and education are associated with higher rates of PC use, as well as living in an urban population center. As can be seen in the tree map in Figure 1, there are vast regional differences in computer use.

Figure 1. Tree map of personal computer use in Russian regions

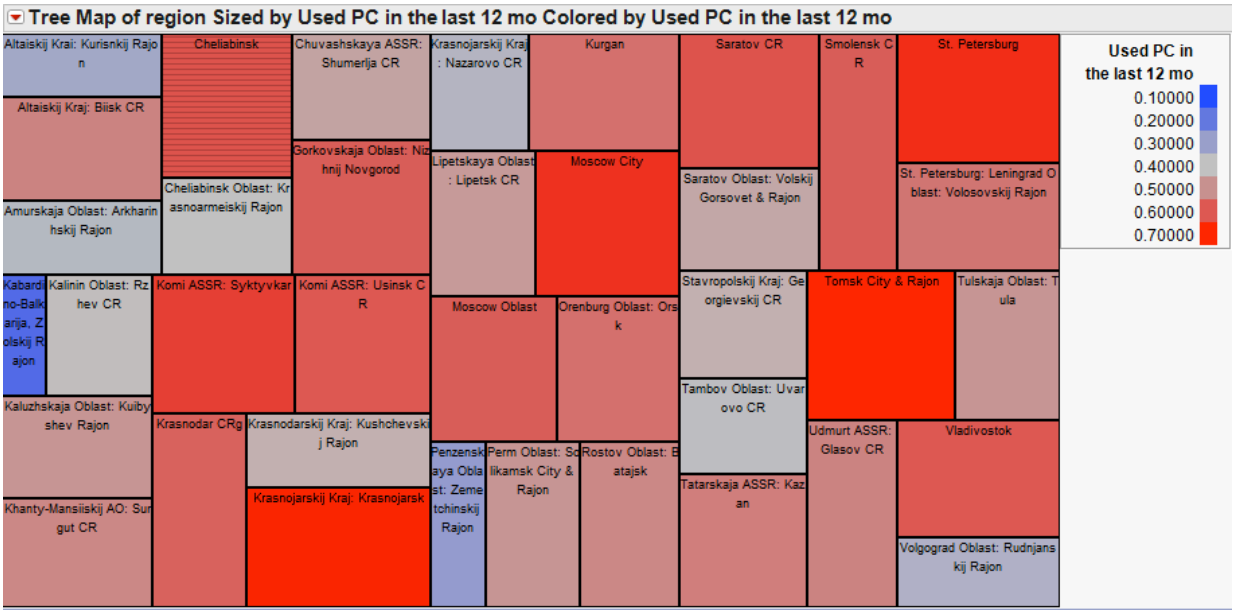
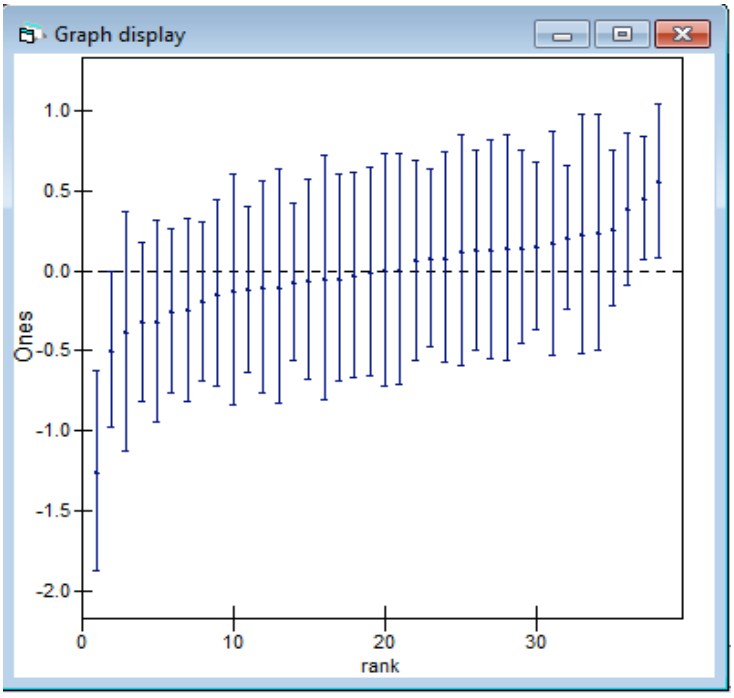


Figure 2. Regional random effects for a multi-level model of PC use in Russia



Interestingly, this difference holds even after accounting for effects of income, education, age and gender. The chart in Figure 2 displays the regional random effects for the 38 regions used in this study, in increasing order, with approximate 95% confidence intervals. The highest positive regional effects are found for St. Petersburg and Moscow Oblast, an expected result. Another region with a high positive effect is Kaluzhskaya Oblast (Kuibyshev Rajon), an unexpected result. Regions with the highest negative effect are Tambov Oblast (Uvarovo CR), Penzenskaya Oblast (Zemetchinskij Rajon) and Amurskaja Oblast (Arkharinhskij Rajon).

References: Story 4

Acilar, A., Markin, M., & Nazarbaeva, E. Exploring the Digital Divide: A Case of Russia and Turkey.

Beketov, H. B. 2009. The role of the digital diversity and the digital divide in Russian development.

Delitsin, L. L. 2006. The problem of digital divide and development potential of the Internet in Russia.

Lihobabin, M. Y. 2006. Gender Determinants of Information Societies.

Norris, P. 2001. *Digital Divide? Civic Engagement, Information Poverty and the Internet in the Democratic Societies*. New York: Cambridge University Press.

Vehovar, V., Sicherl, P., Husing, T., & Dolnicar, V. 2006. Methodological Challenges of Digital Divide Measurements. *The Information Society*, 22(5): 279-290.

A Virtual Academic Living Standards Analytics Community (VALSAC) initiative

Current work on living standards involves researchers in many disciplines (economics, management, sociology, statistics, demography) and many types of institutions (universities, research institutes, statistics offices, Non Governmental Organizations, international organizations, etc.), scattered across scores of countries.

In some respects, researchers have easier access to information than ever, especially through a wealth of resources on the Web. But this deluge of information has costs too, including these:

a. Researchers on issues related to poverty and living standards do not communicate much across disciplines. In part this is because they are already busy keeping up with developments in their own disciplines, and in part because institutions such as universities tend to be organized along disciplinary lines.
b. It is easy for researchers, particularly those starting out, to be overwhelmed by the volume of information. They often need help with technical questions, and guidance in framing questions and keeping a sense of perspective. They also often need the capacity to discern good quality research that may be of use for their needs and purposes given the multitude of research papers available and disseminated over the web.
c. Isolated researchers need encouragement and support. To some extent academic conferences fill this role, but the time and expense involved in gathering large numbers of scattered researchers together in one place are considerable.

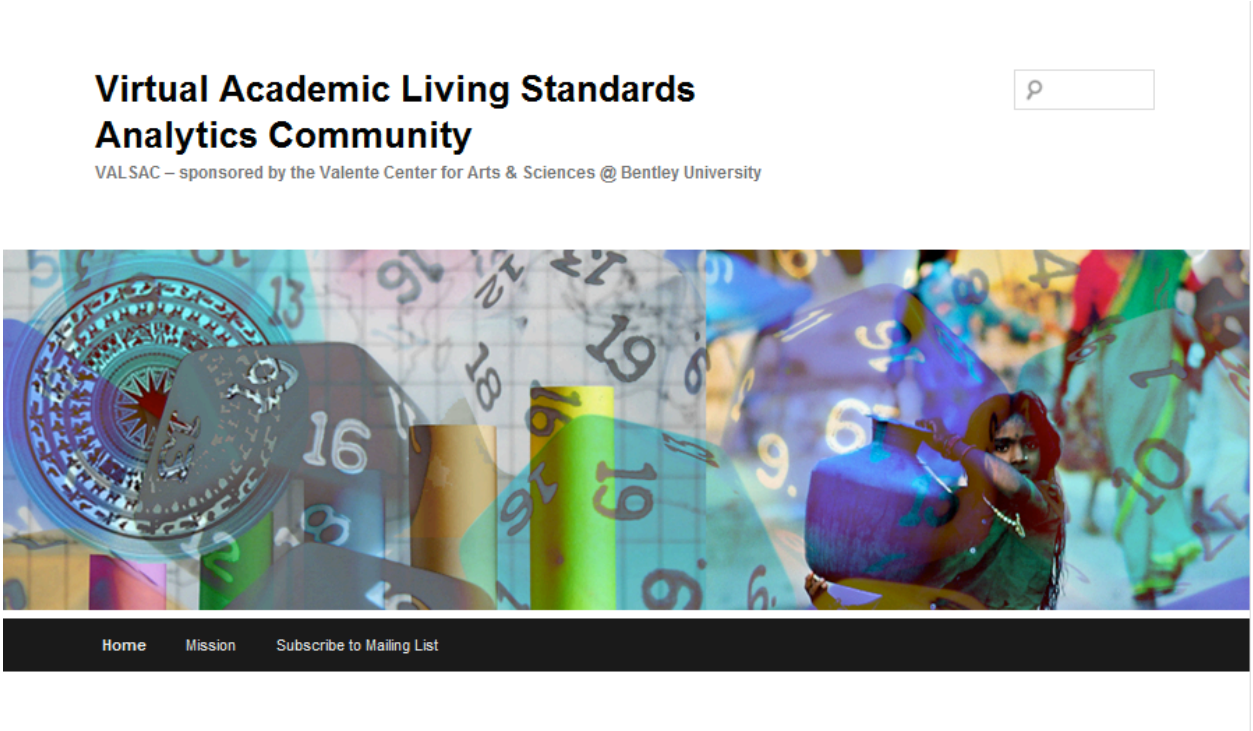
We propose to address these issues by establishing a *Virtual Academic Living Standards Analytics Community (VALSAC)*, where the common thread is the study of living standards using micro-data.

The intention is to complement, not replace, existing resources. Of these, the most important may be the Living Standards Measurement Survey program at the World Bank, which has collected and archived living standards survey data (and associated materials such as questionnaires) from a substantial number of countries. The site has posted 135 working papers related to living standards, typically written with World Bank support or by World Bank Staff, as well as a list of publications (known to LSMS researchers) which make use of living standards data.

VALSAC has two intended central components:

1. A virtual symposium, held about three or four times per year, which anyone with Internet Access, earphones and a microphone, will be able to attend. This will use the synchronous distance education software Centra.
2. A blog, where questions, answers, comments, suggestions and ideas can be posted.

The concept of a Virtual Academic Community is not new; the Global Trade Analysis Project (GTAP), based at Purdue University is an excellent example, although it does not hold a virtual seminar.



To sum up ... a Wordle

To sum up the discussion in this paper, we provide below a "Wordle", a graphical representation of words that tend to occur frequently in the text of this paper (see www.wordle.net on how to construct a worlde on the basis of a body of text). Essentially, the algorithm to build a wordle works by first sorting words in decreasing order of frequency of occurrence in the body of text, and then places words one after the other on the display, in such a way as to avoid intersections of words. A brief description of the algorithm is given in Viégas, Wattenberg and Feinberg (2009).

References: Wordles

Viégas, F., M. Wattenberg and J.Feinberg (2009) 'Participatory visualization with Wordle', available at www.research.ibm.com/visual/papers/worldle_final2.pdf.