centreforcities

Cities Outlook 2009

Technical Appendix: The City Model

Statistical indexes are a popular way to measure performance for a wide range of socio-economic outcomes, from individual to country level. This year we decided to enhance our understanding of city-economies by using formal statistical analysis to rank cities in three different categories: the economic dimension, the society dimension and the built environment dimension. Overall, we found that:

- The **economic** factors that are most able to explain the variation in city performance are those measuring income, employment and skills levels.
- The variables most dominant in the **society** models were those that captured aspects of economic deprivation, such as the proportion of the working population on benefits and Index of Multiple Deprivation scores.
- The **built environment** was more difficult to quantify, largely because the chosen indicators, such as 'per capita carbon emissions' and 'mean house prices', did not have any statistical relationship with each other. Ultimately variables that did not have any statistical 'glue' to other variables had to be excluded, leaving a group dominated by measures of housing quality and price.

This technical Appendix supplements the short appendix provided in the Cities Outlook 2009 publication. It provides a detailed description of the steps taken in collating relevant data and of the statistical techniques and methods used. It explains the statistical units used for cities, missing data treatment, standardisation techniques, index calculation as well as aggregation and weighting methods. This should help make the process by which final ranks were deduced transparent.

The key steps in the process and their aims were:

- **Selecting variables:** To capture aspects of city performance in quantitative variables
- **Standardisation:** To make all variables comparable to each other by transforming them into a normal distribution with a mean of zero and a standard deviation of one
- **Correlation analysis:** To look at relationships between variables and exclude those variables perfectly correlated
- **Principal component analysis:** To further explore the relationship between variables, bundling those related variables and finding what groups of variables explain the most variance in the data
- **Factor analysis:** To develop a formal statistical model based on the results of the principal components analysis. This stage identified the best bundle of variables to be used as formal indicators.
- **Computation of scores:**Factor scores were derived from the regression method and these were used to rank cities. The Economic Prosperity Index and Social Deprivation Index used just one bundle, which meant assigning scores was easy. However, the Built Environment Index used two bundles, each of these had to be weighted according to their explanatory power to derive a final score.

The SPSS computer programme was used for all steps in this analysis.

1. City definitions

As the **geographical unit of analysis** we used CLG's definition of primary urban areas (PUAs). In 2008 we included all English PUAs and Cardiff, Belfast, Edinburgh and Glasgow in *Cities Outlook*. This year we added the other major Scottish and Welsh cities Dundee, Aberdeen, Newport and Swansea.

For the Welsh and Scottish cities we used the corresponding local authority (as recommended by Scottish Enterprise), except for Glasgow, which was defined as an aggregate of the following LAs: West Dunbartonshire, East Dunbartonshire, East Renfrewshire, Renfrewshire and Glasgow city. Belfast was defined as: Belfast City, Carrickfergus, Castlereagh, Lisburn, Newtownabbey and North Down.

However, for some of these cities data was not available for all indicators. This led to the exclusion of Belfast in the built environment ranking and made the inclusion of mean values for some of the cities necessary. Thus, the results for Belfast and Glasgow have to be treated carefully.

We also included countries into our ranking this year, although these do not drive the model. This is helpful in benchmarking cities with regards to national averages.

2. Index Variables

The choice of variables was based on availability, a thorough literature review, and expert judgement. Table 1 summarises the full list of variables considered for inclusion and used in initial stages of statistical analysis. The variables were chosen according to the aim of each of the indexes, specifically:

- Economic model – to measure overall economic development of a city-economy
- Society model – to measure the overall 'health' of a city's society
- Environment model - measure the overall quality of built and the natural environment

For all models there were debates over the definition of what it sought to measure – for instance what constitutes the 'health' of a city's society? It is inevitable that any answer is subjective. In the absence of precise definitions we had to rely on background literature and common sense.

Table 1: The initial set of variables considered for inclusion in the indexes

Economic Model

- % of people employed in public administration, education and health 2006 (a measure of a city's resilience)
- % of employees employed in private services 2006 (a measure of an economy's health and an aggregate of the % employed in: Distribution, Hotels and Restaurants; Transport and Communications; Banking, Finance and Insurance; Other Services)
- % of economically active within a city with no qualifications 2007
- % of economically active within a city with NVQ Level 4 qualifications 2007
- % of employees that are highly skilled 2007 (the percentage of those employed as managers and administrators, those employed in professional occupations, and those in associated professional and technical occupations; and indicator of the supply side of the labour market)
- the employment rate 2007
- the average income in a city 2005-06
- the stock of VAT registered businesses per 10,000 adult population 2006 (an indicator for enterprise and innovation activity)
- GVA per capita 2005
- weekly gross pay 2007
- mean tax take 2007

Society Model

- first docile resident income 2006
- % of working age population on incapacity benefit 2007
- % of working age population on income support 2007
- % of working age population on job-seekers allowance 2007
- an aggregate of the % of people on these three kinds of benefits 2007
- the male life expectancy 2003-05 (rolling average)
- the female life expectancy 2003-05 (rolling average)
- robberies per 10,00 population 2006-07
- the median Index of Multiple Deprivation (IMD) score 2007
- the minimum IMD score 2007
- the maximum IMD score 2007
- the IMD range 2007

Environment Model

- per capita carbon emissions 2006
- % of people commuting to work by car 2001
- Rateable value per m2 (RV) all bulk classes 2007
- RV retail premises 2007
- RV offices 2007
- RV factories 2007
- RV warehouses 2007
- % of dwelling stock in Council tax bands A+B 2006
- % of dwelling stock in Council tax bands G+H 2006
- LA dwellings as % of total stock 2006
- Registered Social Landlord (RSL) as % of total stock 2006
- Owner occupied dwellings as a % of total stock 2006
- Total unfit dwellings as a % of total stock 2006
- Mean house prices 2007
- Affordability ratio (mean house prices/annual average earnings) 2007
- Basket of retail establishments per 10,000 population 2006 (includes retail; hotels, restaurants and bars; recreational, cultural and sports establishments; beauty and wellbeing establishments

3. Standardisation of variables

Due to the different scales used in each variable there was a process of standardisation. This ensures that each variable has the same weighting in the classification. If left un-standardised certain variables, such as average income would dwarf the importance of those indicators measured as a percentage because of the higher values and larger range the income data is stretched over. It would also create outliers based solely on the average income variable. Thus, if left unstandardised, variables used in statistical analysis would add bias to the dataset.

To eradicate this bias all variables were standardised using z-scores. This is a common way to control for variables with different means and standard deviations (see Box 1). Z-scores can be used when variables are normally distributed and there are no extreme outliers. Those variables without a normal distribution had to first go through a log transformation, this helps to correct for skew by squashing the right tail of the distributions.[1]

1. Social and economic data does tend to have a positive skew because extreme outliers mean that the majority of areas have a value below the average.

All variables were plotted to check for skew and extreme outliers – for the majority there was a skew – and hence nearly all variables had to be first put through a log transformation and then a z-score transformation.

Box 1: Z-scores and Log transformations

Z-scores are a way to transform data variables with different means and/or standard deviations. The z score for an item, indicates how far and in what direction, that item deviates from its distribution's mean, expressed in units of its distribution's standard deviation. Such that:

$$Z_x = \frac{X - \mu_x}{\sigma_x}$$

Where:
X = the value
μ_x = the mean of the series
σ_x = the standard deviation of the series

The mathematics of the z-score transformation are such that if every item in a distribution is converted to its z-score, the transformed scores will necessarily have a mean of zero and a standard deviation of one.

Log transformations
Log transformations are used to correct for positive skew as it squashes the right side of the distribution. Log transformations cannot be performed on the value of zeros or on negative numbers, in that case a constant, e.g. lox (X_i+1) must be added.

4. Calculation of indices

Once variables were standardised the statistical techniques to build indices could be applied.

Stage 1 - Correlation analysis

A simple correlation analysis was conducted for all the bundles of indicators. Strong correlations within a dataset are undesirable, as they represent data redundancy. Highly correlated variables mean the bulk of information is contained within just one of the variables, and having both variables just repeats information. Including highly correlated variables makes it very hard to gauge the effect of any individual variable and can lead to double counting. Tables 2-4 detail the correlation between variables for each of the models.

A number of strong correlations were found in all three models. Those with a correlation of higher then 0.7 represent a significant redundancy. For each of the variables the correlations are not always surprising – for instance the Economic variables "PUA Income (£) 2005-06" and "Mean tax take (£) 2005-06" are strongly correlated. This is not surprising because the mean tax take should be closely related to PUA income.

Therefore, there is no reason to include "mean tax take." Similarly, weekly pay gross and PUA income are highly correlated with r=0.93 and seem to contain essentially the same information. Overall, correlations can be split into three types:

- The first are pairs of variables which share the same denominator, so that the correlations will have a natural propensity to be negative. For example the percentage of employees in public administration, education, and health and the percentage of employees in the private services are slices of the same 100% pie, they are inversely correlated. This will produce a contrast (i.e. a component) in the principal components analysis in the Economic Model. In this case, as long as one of the variables that make up the pie is excluded, others can be left in.
- The second type of correlation consists of those variables that are inherently connected due to causality i.e. one is fundamentally a property of the other, but they don't share the same denominator. An example of this is the correlation between the maximum IMD score and the range of IMD scores is r=0.99 in the Society model. It might therefore be necessary to exclude either the minimum and maximum IMD score variables or the IMD range variable from the factor model.
- The third type of correlation is made up of correlations between variables where the presence of one indicates the presence or absence of another but does not fundamentally cause it to be so. For example, a city with a high percentage of population with no skills and the PUA income.

Each of these types of correlations has different implications for the next stage of analysis. For instance, variables that repeat a lot of the same information will be taken out, such as mean tax take. The Environmental model has some very high correlations between some of the variables (e.g. 'mean house prices' and 'RV all bulk class' or 'mean house prices' and 'affordability ratio') whilst other variables are effectively uncorrelated (e.g. Proportion of unfit dwellings and proportion of owner occupier dwellings), and not all variables in the dataset are positively correlated. This suggests that the variables in the dataset are indicators of distinct aspects of the environmental quality that might need to be treated as separate building blocks in the overall city index. This point will be further explored using Principal Component Analysis.

Table 2: Correlations between the variables for the Economic Model

	Public admin, edu, health employees	Private services employees aggregate	Eco active with no qual %	Highly Skilled %	Employment rate	PUA Income (£) 2005-06	Mean tax take (£) 2005-06	VAT Stock	GVA Per Capita (£)	Weekly Pay Gross
Public admin, edu, health employees	1.00									
Private services employees aggregate	-0.66	1.00								
Eco active with no qual %	-0.01	-0.11	1.00							
Highly Skilled %	-0.01	0.37	-0.54	1.00						
Employment rate	-0.00	0.10	-0.37	0.32	1.00					
PUA Income (£) 2005-06	-0.14	0.39	-0.48	0.82	0.44	1.00				
Mean tax take (£) 2005-06	-0.14	0.39	-0.48	0.80	0.41	**1.00**	1.00			
VAT Stock	-0.10	0.24	-0.25	0.62	0.51	0.78	0.77	1.00		
GVA Per Capita (£)	0.03	0.11	-0.37	0.58	0.40	0.58	0.58	0.42	1.00	
Weekly Pay Gross	-0.19	0.39	-0.42	0.75	0.47	**0.93**	**0.93**	0.76	0.52	1.00

Table 3: Correlations between the variables for the Society model

	Residents 10 Percentile	Working age on Incapacity benefit %	Working age on Income Support %	Working age on JSA %	Working age on benefits(IB+IS+JSA)%	Male Life Expectancy	Female Life Expectancy	Robberies	IMD median score 2007	IMD min score 2007	IMD max score 2007	IMD range 2007
Residents 10 Percentile	1.00											
Working age on Incapacity benefit %	-0.70	1.00										
Working age on Income Support %	-0.42	0.35	1.00									
Working age on JSA %	-0.52	0.42	0.87	1.00								
Working age on benefits %	-0.72	0.86	0.76	0.82	1.00							
Male Life Expectancy	0.73	-0.81	-0.62	-0.72	-0.90	1.00						
Female Life Expectancy	0.61	-0.77	-0.56	-0.60	-0.82	0.90	1.00					
Robberies	0.19	-0.26	0.45	0.42	0.09	-0.13	-0.07	1.00				
IMD median score 2007	-0.66	0.70	0.79	0.82	0.91	-0.85	-0.79	0.24	1.00			
IMD min score 2007	-0.60	0.58	0.49	0.48	0.64	-0.60	-0.60	-0.05	0.69	1.00		
IMD max score 2007	-0.62	0.68	0.62	0.64	0.80	-0.77	-0.65	0.17	0.72	0.37	1.00	
IMD range 2007	-0.55	0.62	0.57	0.60	0.73	-0.70	-0.58	0.19	0.64	0.22	**0.99**	1.00

Table 4: Correlations between the variables for the Environmental model

	Per Capita Carbon Emissions	Car communter %	RV All Bulk Classes	RV Retail Premises	RV Offices	RV factories	RV warehouses	Council Tax Bands A+B	Council Tax Bands G+H	LA Dwelling Proportion	RSL Dwelling Proportion	Owner occupier Dwelling Proportion	Total Unfit Dwellings Proportion	Mean House Prices	Affordability ratio	Basket of Retail Establishments
Per Capita Carbon Emissions	1.00															
Car communter %	0.24	1.00														
RV All Bulk Classes	-0.19	-0.62	1.00													
RV Retail Premises	-0.08	-0.51	0.83	1.00												
RV Offices	-0.14	-0.50	0.88	0.81	1.00											
RV factories	-0.15	-0.29	0.87	0.64	0.71	1.00										
RV warehouses	-0.18	-0.27	0.86	0.66	0.78	0.93	1.00									
Council Tax Bands A+B	0.13	0.29	-0.85	-0.61	-0.72	-0.88	-0.86	1.00								
Council Tax Bands G+H	-0.03	-0.33	0.81	0.61	0.77	0.77	0.79	-0.85	1.00							
LA Dwelling Proportion	-0.12	-0.27	0.08	0.25	0.16	-0.10	-0.03	0.18	-0.09	1.00						
RSL Dwelling Proportion	0.22	0.11	-0.21	-0.18	-0.14	-0.17	-0.19	0.22	-0.07	-0.71	1.00					
Owner occupier Dwelling Proportion	-0.11	0.22	0.12	-0.15	-0.07	0.32	0.25	-0.48	0.16	-0.49	-0.26	1.00				
Total Unfit Dwellings Proportion	-0.05	-0.14	-0.21	-0.27	-0.26	-0.28	-0.33	0.29	-0.19	-0.20	0.24	-0.00	1.00			
Mean House Prices	-0.18	-0.51	0.94	0.71	0.82	0.85	0.84	-0.93	0.86	-0.05	-0.21	0.29	-0.26	1.00		
Affordability ratio	-0.24	-0.53	0.82	0.59	0.64	0.71	0.68	-0.81	0.63	-0.09	-0.21	0.35	-0.20	0.90	1.00	
Basket of Retail Establishments	-0.20	-0.50	0.54	0.38	0.30	0.36	0.28	-0.48	0.36	-0.05	-0.19	0.29	0.03	0.55	0.66	1.00

Stage 2 – Principal Component Analysis

A principle component analysis (PCA) was conducted on all variables, excluding the redundant variables identified in the correlation analysis (see Box 2). By detecting structures in the relationships between variables, PCA analysis bundles correlated variables and helps to condense the number of variables (see Box 3). It reveals how much of the variance in the data can be explained by each of the bundled factors (through eigenvalues) and was used to identify the factors to be used in the final model.

A separate principal components analysis was performed on the set of environment, society and economic variables. Tables 5-10 show the main results; these are the component loadings, eigenvalues and a statistically optimised solution (rotated components). The component loadings reported in the columns show the contribution each of the individual variables makes to the respective components. The variables with the highest values carry the most "weight" and are the strongest indicators for the respective component.[2]

The results suggest that the economic, society and environment model variables in the dataset can be adequately summarised in one component for each of these three areas. Beyond the "weighted" summary of the variables in the first component, we can get additional information about the underlying structure by inspecting the second component.

- **Economic model** – The first component accounts for more then half of the total variance and is dominated by measures of overall economic development, such as having a highly skilled population. The component matrix shows that the second component is made up of public sector employment and private services employment, which is why they are inversely correlated. This second bundle is picking up on the differences in the structure of the labour market of the cities.

- **Society model** – The first component accounts for over 67 percent of the total variance, with the second component only picking up on robberies. The first bundle is measuring aspects of social deprivation, hence the bundling of claimant counts and life expectancy. Again there are very high correlations between IMD derived variables and some of these will have to be taken out for the next stage in the analysis to avoid double counting.

- **Environment model** – This model went through several iterations before the select variables were chosen. As shown in the correlation analysis, the disconnection between the variables such as carbon emissions and other variables meant that it was sensible to exclude such variables from the index. The final PCA output for this model shows that first component is made up of variables measuring housing value, cost and local retail amenities. It accounts for about a half of total variance, but the second component, which picks up on aspects of quality of housing adds a further 20 percent to the variance explained.

In summary, the PCA results suggest that set of economic and societal variables can be summarised into one component each, although the environmental model probably needs a more differentiated consideration. The addition of the second component captures the vast majority of the variance (see scree plots, Figure 1-3) so that one can be confident that these variables are reflecting the underlying difference between PUAs.

2. Those bundles with eigenvalues of less then one were ignored in line with Kaiser test (see Box 2).

PCA for Economic Model

Table 5: Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings	
Component	Total	% of Variance	Cumulative %	Total	Cumulative %
1	4.53	50.41	50.41	4.54	50.41
2	1.62	18.06	68.47	1.63	68.47
3	0.85	9.47	77.93		
4	0.67	7.40	85.33		
5	0.52	5.82	91.16		
6	0.35	3.85	95.01		
7	0.24	2.68	97.69		
8	0.14	1.59	99.28		
9	0.06	0.72	100.00		

Extraction Method: Principal Component Analysis.

Figure 1: Scree plot for the Economic model

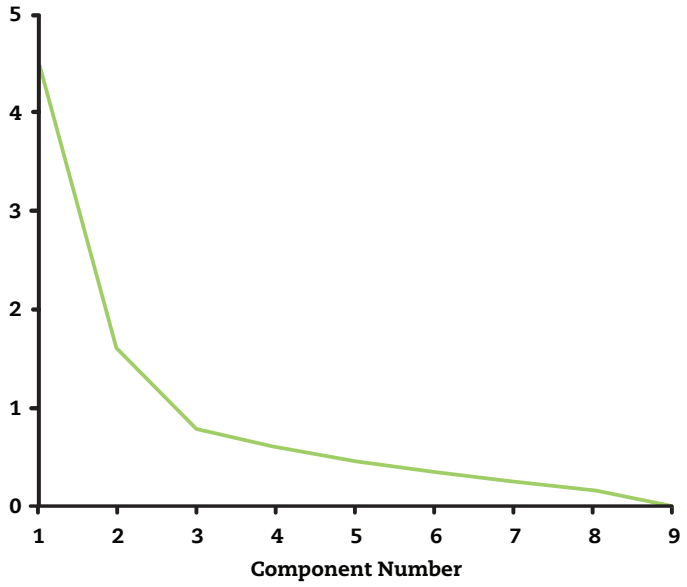

Table 6: Rotated Component Matrix

	Component	
	1	2
Public admin, education, health employees 2006	0.07	**-0.92**
Private services employees aggregate 2006	0.22	**0.89**
Economically active with no qualifications 2007	**-0.64**	-0.09
Highly Skilled 2007	**0.87**	0.17
Employment rate 2007	**0.63**	-0.04
PUA income 2005-06	**0.91**	0.25
VAT stock 2006	**0.77**	0.15
GVA per capita 2005	**0.75**	-0.03
Weekly pay gross 2007	**0.86**	0.30

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalisation. Rotation converged in 3 iterations.

PCA for Society Model

Table 7: Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings	
Component	Total	% of Variance	Cumulative %	Total	Cumulative %
1	8.06	67.20	67.20	8.06	67.20
2	1.62	13.47	80.67	1.62	80.67
3	0.92	7.67	88.34		
4	0.45	3.78	92.12		
5	0.33	2.73	94.84		
6	0.24	2.01	96.85		
7	0.16	1.29	98.14		
8	0.10	0.87	99.01		
9	0.07	0.60	99.61		
10	0.04	0.36	99.97		
11	0.00	0.01	99.99		
12	0.00	0.01	100.00		

Extraction Method: Principal Component Analysis.

Figure 2: Scree plot for Society model

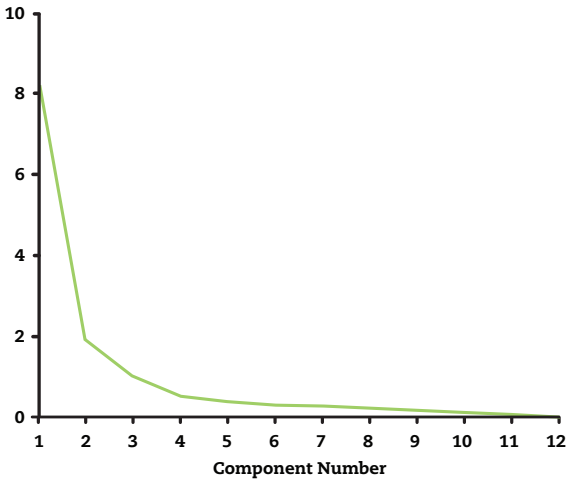

Table 8: Rotated Component Matrix

	Component	
	1	2
Residents 10 percentile 2006	**-0.79**	0.39
Working age on benefits 2007	**0.86**	-0.43
Working age population on IB 2007	**0.79**	0.43
Working age population on IS 2007	**0.86**	0.37
Working age population on JSA 2007	**0.98**	-0.03
Robberies 2006-07	0.18	**0.91**
IMD median 2007	**0.94**	0.10
IMD min score 2007	**0.69**	-0.28
Male Life Expectancy 2003-05	**-0.93**	0.06
Female Life Expectancy 2003-05	**-0.84**	0.12
IMD max score 2007	**0.87**	0.07
IMD range 2007	**0.81**	0.11

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalisation. a 2 components extracted.

PCA for the Environmental Model

Table 9: Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings	
Component	Total	% of Variance	Cumulative %	Total	Cumulative %
1	3.72	53.08	53.08	3.72	53.08
2	1.19	17.00	70.08	1.19	70.08
3	0.93	13.23	83.34		
4	0.54	7.72	91.06		
5	0.43	6.18	97.23		
6	0.10	1.47	98.70		
7	0.09	1.30	100.00		

Extraction Method: Principal Component Analysis.

Figure 3: Scree Plot for the Environment model

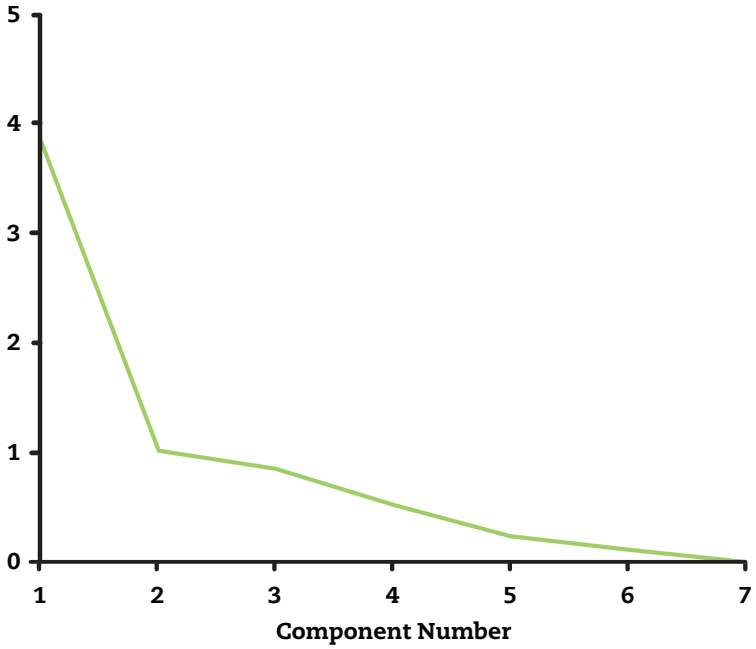

Table 10: Rotated Component Matrix

	Component	
	1	2
All bulk classes 2007	**0.85**	0.26
Council tax bands G-H 2006	**0.85**	-0.11
Council Tax Bands A+B 2006	**-0.94**	-0.02
Mean house prices 2007	**0.93**	0.16
Basket of retail establishments 2006	**0.68**	-**0**.14
Total unfit dwellings 2006	-0.16	**-0.71**
LA Dwelling Proportion of total 2006	-0.13	**0.76**

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalisation. a Rotation converged in 3 iterations.

Box 2: Principal Component Analysis - revealing the relationships between variables

PCA is recommended as an exploratory tool to uncover unknown trends in the data. It is a method that reduces data dimensionality by performing a covariance analysis between factors. Covariance is always measured between two factors. So with three factors, covariance is measured between factor *x* and y; y and z, and *x* and z. When more than two factors are involved, covariance values can be placed into a matrix. This process deduces the linear combinations of original variables.

In other word, the *p* original variables are combined into *q* linear combinations, which form the new principal components of the system. A standardized linear combination Z_1 of a data vector, $X_1=(X_{11}, X_{12}, \ldots, X_{1p})$, of length *p* is defined as:

$Z_1=w_{1t}X_1$, where the sum of the squares of the weights, w_i, is 1.

One key decision when carrying out PCA analysis is the rotation method. The goal of rotation is to simplify and clarify the data structure. Rotation cannot improve the basic aspects of the analysis, such as the amount of variance extracted from the items. As with extraction method, there are a variety of choices. Varimax rotation is by far the most common choice. Varimax, quartimax, and equamax are commonly available orthogonal methods of rotation; direct oblimin, quartimin, and promax are oblique. Orthogonal rotations produce factors that are uncorrelated; oblique methods allow the factors to correlate.

PCA will find Eigenvectors and eigenvalues relevant to the data using a covariance matrix. Eigenvectors can be thought of as "preferential directions" of a data set, or in other words, main patterns in the data. For either PCA, there cannot be more components than there are conditions in the data. Eigenvalues can be thought of as quantitative assessment of how much a component represents the data. The higher the eigenvalues of a component, the more representative it is of the data. Eigenvalues can also be representative of the level of explained variance as a percentage of total variance. By themselves, eigenvalues by are not informative. The percent of variance explained is dependent on how well all the components summarize the data. In theory, the sum of all components explains 100% variability in the data.

There are two ways to decide how many bundles to use:

1. **The Kaiser criterion** - only factors with eigenvalues greater than one are retained. In essence this is like saying that, unless a factor extracts at least as much as the equivalent of one original variable, it is not worth keeping in.
2. **The scree test** - a graphical method is the scree test, where eigenvalues are plotted in a simple line plot and the sharp bend in the line or the 'elbow' is taken to indicate the number of bundles that explain the majority of the variance.

In practice it is a combination of these two that is used to make decisions on the number of components to be used.

Stage 3 – Factor Analysis

Factor analysis is a variable reduction technique, similar to PCA. The major difference between the two is that whilst PCA is a descriptive technique, a factor analysis is a formal statistical model (see Box 3). Unlike in PCA, the factor model tries to account for all the variance in the data using one factor and enables formal assessment of how well this one factor describes the data through a goodness-of-fit significance test. These factor models can be used to calculate factors scores for PUAs, so that they can be ranked, as will be discussed in the next section.

For each of the models the key results are:

- **Economic model** – The one factor model has a high goodness of fit (see Table 14). The first factor explains more then half of the variance in the data (i.e. the differences between cities). The variables that contribute most to the explanation of economic performance are those measuring PUA income and weekly gross pay. Variables that capture the structure of the labour market, for example, the percentage of private service employees are weaker indicators and thus carry less weight (smaller factor loadings).

- **Society model** – This model went through several iterations before the final model was decided on. The variables in the original dataset captured a wide range of aspects that could not be easily reduced into one indicator. Thus, three different model specifications were tested in order to find a model that gives an adequate summary of the variables and that does justice to the data (i.e. passes the goodness of fit test), these included:

  1. The **aggregate model** – This model used a balanced choice of indicators, meaning it avoided including indicators that essentially said the same thing. For instance it excluded the separate variables on JSA, IB and IS, and included the combined variable.

  2. The **disaggregated model** – This did use the maximum number of variables, including those that were highly correlated.

  3. The **neat model** – Finally, this model included only variables of multiple deprivation and benefit dependency, excluding measures of citizen well-being. The aim of this approach was to minimise the contrast prevalent between indicators.

  The first model proved to be best able in explaining the differences in the health of a city's society. The first factor accounts for over half of the variance and the factor loadings are generally high and positive. 'Robberies' does have a low factor loading but the goodness of fit is still acceptable.

- **Environment model** – This model proved more difficult to interpret and had a poor goodness of fit outcome. The first factor accounts for less than half of the variance and there are a number of low and negative loadings. Based on the cumulative evidence of the correlation, PCA and factor analysis it was decided that two bundles of indicators would be needed (see final variables in next section).

Economic Model

Table 11: Communalities[3]

	Initial	Extraction
Private services employees aggregate 2006	0.21	0.16
Highly skilled 2007	0.71	0.69
Employment rate 2007	0.31	0.23
PUA income 2005-06	0.89	0.97
VAT stock 2006	0.56	0.54
GVA per Capita 2005	0.45	0.37
Weekly gross pay 2007	0.85	0.87

Extraction Method: Maximum Likelihood.[4]

Table 12: Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.17	59.54	59.54	3.82	54.60	54.60
2	0.97	13.83	73.38			
3	0.67	9.54	82.92			
4	0.59	8.45	91.37			
5	0.31	4.49	95.85			
6	0.22	3.10	98.95			
7	0.07	1.05	100.00			

Extraction Method: Maximum Likelihood.

Table 13: Factor Matrix

	Factor
	1
Private services employees aggregate 2006	0.38
Highly skilled 2007	0.83
Employment rate 2007	0.48
PUA income 2005-06	0.99
VAT stock 2006	0.74
GVA per Capita 2005	0.61
Weekly gross pay 2007	0.93

Extraction Method: Maximum Likelihood.

Table 14: Goodness-of-fit Test

Chi-Square	df	Sig.
17.30	14	0.24

3. Communalities are the proportion of a variable's variance explained by a factor structure
4. The Maximum Likelihood is a method of parameter estimation in which a parameter is estimated to be that value for which the data are most likely.

Society Model

Table 15: Communalities

	Initial	Extraction
Residents 10 percentile 2006	0.63	0.59
Working age on JSA 2007	0.77	0.85
Robberies 2006-07	0.27	0.01
IMD median score 2007	0.77	0.79
Male Life Expectancy 2003-05	0.54	0.57
IMD range 2007	0.52	0.55

Extraction Method: Maximum Likelihood.

Table 16: Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.67	61.20	61.20	3.37	56.10	56.10
2	1.10	18.34	79.55			
3	0.46	7.68	87.22			
4	0.37	6.21	93.43			
5	0.25	4.09	97.52			
6	0.15	2.48	100.00			

Extraction Method: Maximum Likelihood.

Table 17: Factor Matrix

	Factor 1
Residents 10 percentile 2006	-0.77
Working age on JSA 2007	0.92
Robberies 2006-07	0.12
IMD median score 2007	0.89
Male Life Expectancy 2003-05	-0.76
IMD range 2007	0.74

Extraction Method: Maximum Likelihood.

Table 18: Goodness-of-fit Test

Chi-Square	df	Sig.
16.98	9	0.05

Environment model

Table 19: Communalities

	Initial	Extraction
All bulk classes 2007	0.80	0.76
Council tax bands G-H 2006	0.75	0.60
Council tax bands A+B 2006	0.87	0.84
Mean house prices 2007	0.85	0.89
Basket of retail establishments 2006	0.34	0.32
Total unfit dwellings 2006	0.13	0.04
LA Dwelling Proportion of total 2006	0.15	0.00

Extraction Method: Maximum Likelihood.

Table 20: Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.72	53.08	53.08	3.45	49.35	49.35
2	1.19	17.00	70.08			
3	0.93	13.26	83.34			
4	0.54	7.72	91.06			
5	0.43	6.18	97.23			
6	0.10	1.47	98.67			
7	0.09	1.30	100.00			

Extraction Method: Maximum Likelihood.

Table 21: Factor Matrix

	Factor 1
All bulk classes 2007	0.87
Council tax bands G-H 2006	0.78
Council tax bands A+B 2006	-0.21
Mean house prices 2007	0.94
Basket of retail establishments 2006	0.56
Total unfit dwellings 2006	-0.91
LA Dwelling Proportion of total 2006	-0.05

Extraction Method: Maximum Likelihood.

Table 22: Goodness-of-fit Test

Chi-Square	df	Sig.
44.26	14	0.00

Box 3: Factor Analysis – a formal statistical model to produce final indicators

Factor Analysis:

- Is a variable reduction technique which identifies the number of latent constructs and the underlying factor
- Hypothesizes an underlying construct, a variable not measured directly
- Estimates factors which influence responses on observed variables
- Allows you to describe and identify the number of latent constructs (factors)

The aim of factor analysis is to reveal any latent variables that cause the manifest variables to covary. During factor extraction the shared variance of a variable is partitioned from its unique variance and error variance to reveal the underlying factor structure; only shared variance appears in the solution. Principal components analysis does not discriminate between shared and unique variance.

Thus: $Y = X\beta + E$

where Y is a matrix of measured variables
X is a matrix of common factors
β is a matrix of weights (factor loadings)
E is a matrix of unique factors, error variation

The amount of variance explained is the trace (sum of the diagonals) of the decomposed adjusted correlation matrix. Eigenvalues indicate the amount of variance explained by each factor. Eigenvectors are the weights that could be used to calculate factor scores. In common practice, factor scores are calculated with a mean or sum of measured variables that "load" on a factor.

Scores and weighting

Finally, factor scores were assigned based on the formal statistical models highlighted above, using the regression method. This gave cities scores based primarily on those factors that were most important in explaining the variance for each model. Cities that had missing values on some of the variables had to be given mean values to fill these gaps.

For the environmental model we chose to have two bundles. Each of these were assigned factor scores and then weighted according to their explanatory power. The different bundles were not given equal weight because this would ignore the extent to which each bundle was correlated with the underlying factor. The first bundle explained 53 percent, and the second 17 percent of the variance. Thus, taken together, the first bundle was given a weight of 0.76, and the second 0.24.

Table 23: Final Variables

The Economic Development Index

- PUA income – the average income in a city 2007
- Gross weekly pay 2007
- GVA per capita 2005
- VAT stock per 10,000 adult population 2006 (an indicator for enterprise and innovation activity)
- Employment rate 2007
- % of highly skilled workers 2007 (those employed as managers and senior administrators, in professional occupations, and in associated professional and technical occupations)
- % of employees in private services 2006 (the aggregate of those employed in distribution; hotels and restaurants; transport and communication; banking, finance and insurance; and other services)

The Social Deprivation Index

- the weekly wage threshold below which 10% of the working population fall 2006
- the Index of Multiple Deprivation (IMD) range 2007 (the range between maximum and minimum IMD scores)
- male life expectancy 2003-05 (rolling average)
- robberies per 10,000 adult population 2006-07
- the percentage of the working age population on either or several of three types of benefits 2007 (incapacity benefit, income support and jobseeker's allowance)
- the IMD median score 2007

The Environment Model

Bundle 1

- the rateable value per square metre of all bulk classes (non-residential) property 2007
- the percentage of dwellings in council tax bands G & H 2006 (houses in the two highest valuation categories for council tax purposes)
- mean house prices 2007
- retail establishments per 10,000 inhabitants 2006 (a basket composed of retail; hotels; restaurants and bars: recreational; beauty and well-bring establishments)

Bundle 2

- the percentage of dwellings in council tax bands A & B 2006 (houses in the two lowest valuation categories for council tax purposes)
- total unfit dwellings as a percentage of total housing stock 2006
- Local Authority housing as a percentage of total stock 2006

Overall the number of variables were reduced from ten to seven, 12 to six, 16 to seven for the economic, society and built nvironment models respectively, with the methodological process minimising double counting and maximising the reliability of the final ranks.

The Centre plans to continue to improve the Index through a wider search of appropriate indicators and development of the methodology for next publication of Cities Outlook in 2010.

Acknowledgements

This appendix uses analysis and the subsequent write-up provided by Katrin Hohl (freelance analyst).

centreforcities

Enterprise House
59 - 65 Upper Ground
London SE1 9PQ

t **020 7803 4300**
www.centreforcities.org

March 2009