The Governance Report

# Indicator Dashboards Methodological Notes

Piero Stanig*

March 1, 2013

In this short note, we clarify some methodological choices made in the preparation of the Governance Report Indicators Dashboards. The purpose of this document is to allow the academic research user to understand and evaluate the choices that were made when assemblying the data. It does not replace the fuller discussion of the indicators project found in Hertie School (2013a; 2013b). In particular, users are referred to Anheier et al. (2013), Stanig (2013), and Stanig and Kayser (2013) for a relatively detailed outline of the conceptual grounding of the indicators project. It should be kept in mind that, as of 2013, the Indicator Dashboards are in their first edition, and we consider them a stepping stone towards a fuller governance measurement system. For this reason, questions, comments, and constructive criticism are more than welcome.

## 1 Dashboards

We advocate for the presentation of indicators in the form of "dashboards", i.e. sets of key indicators related to a broader concept of interest, rather than producing one (or many) aggregate indexes. The dashboard approach provides information that can be used to create aggregate indexes that fit the purpose of the specific analysis one is interested in. Moreover, it avoids the problem of conflating on a single dimension inherently

*Research Fellow in Governance and Methodology, Hertie School of Government. stanig@hertie-school.org

multi-dimensional phenomena. Users are referred to Stanig and Kayser (2013) for some discussion of these issues. We do present a limited set of mid-range aggregate indexes that, we believe, have heuristic value and make it possible to detect interesting patterns in the data. These should not be treated as the "ultimate" measures of the phenomena we investigate.

There are three dashboards: the Transnational Governance Dashboard, the National Governance Dashboard, and the City Governance Dashboard. The codebooks with variable definitions, and the dashboards themselves, in Stata, CSV, and Excel formats, can be downloaded from `www.governancereport.org`.

## 2 National Governance Dashboard

The national dashboard captures three key dimensions of governance: internal governmental implementation capacity (Effectiveness), external expertise (Efficacy), and strength of civil society.

### 2.1 Effectiveness

Effectiveness is measured as

1. the Weberianness of a bureaucracy, i.e. its impartiality and professionalism in hiring and promotion;
2. the statistical capacity, i.e. the ability to diagnose problems through the collection of social and economic data;
3. the intellectual resources within government, measured by the number of researchers with an advanced degree employed by the government.

   We measure (1) with data from the Quality of Government (QoG) Institute (Teorell et al. 2011) that address the issue of Weberianness with a survey of experts in a large set of countries. The survey was carried out between 2008 and 2010. The authors of the report summarize the answers to several questions in three indices, of which two pertain for our purposes: the index of bureaucratic impartiality and the index of bureaucratic professionalism. The index of impartiality measures to what extent government institutions exercise their power impartially. The impartiality norm is defined as follows: "When implementing laws and policies, government officials shall not take into consideration anything about

the citizen/case that is not beforehand stipulated in the policy or the law" (Rothstein and Teorell 2008: 170). The index of professionalism measures to what extent the public administration is professional rather than politicized (Dahlström et al. 2011). Higher values indicate a more professionalized public administration.

The index of statistical capacity compiled by the World Bank for over 140 developing countries. [1] Using information available from the International Monetary Fund, United Nations, UNESCO, and the World Health Organisation and its own information, the World Bank scores a country against specific criteria along three dimensions (statistical methodology, source data, and periodicity and timeliness) and derives an overall score for each country on a scale of 0-100, with a score of 100 indicating that the country meets all the criteria.

We measure Intellectual resources within the public administration with a variable collected for several countries by UNESCO : the (log) full-time equivalent (FTE) number of holders of advanced degrees (in all fields) employed by the government.

The four variables are aggregated (via simple average of the standardized and centered sources) into an "effectiveness" index. Given that the pattern of missingness of the statistical capacity index at the time of the Governance Report indicators preparation (last months of 2011) is related to the level of development of the country (namely, the vast majority of advanced countries were not scored), it would be misleading to just center the variable before aggregation, without imputing the missing values. Indeed, due to the observed pattern of missingness, the average level of statistical capacity (which would be assigned a score of 0 in the standardization and centering step) refers to a country which is less developed than the average country in our collection. In order to address this problem, we first impute the missing statistical capacity score based on a regression of the observed scores on log GDP and an indicator variable for countries that are classified as "High Income: OECD". The imputed values are used exclusively to calculate the mean and the standard deviation of the variable in the scaling step: they do not enter the estimation of the index. [2]

[1] Possible minor discrepancies between the values we use and those reported by the source at later times are due to updating of the figures after the Fall of 2011, when we prepared the data.

[2] Not that we have anything *against* imputation, especially proper multiple imputation *a*-la-Rubin (1987; 1996), which would be the optimal way of dealing with *all* the missing values in our data. But we know that governance index users are often a bit wary of

Similarly, the UNESCO figures display a pattern of missingness related to development. Hence before computing mean and standard deviation for the purpose of rescaling, we impute the values based on a regression on log GDP, the "High Income: OECD" indicator, and log population. Again, the imputed values are used only in the scaling step, not the index estimation step.

Standard errors for the index are estimated in the following way. First of all, we build a rough estimate of the standard error of the professionalism and impartiality indexes, by dividing the confidence range (estimated by the Quality of Government Institute via bootstrap) by four. The UNESCO and WB figures do not come with an estimate of variability. To estimate the uncertainty associated with the estimates of this index, we use an approach which is somewhat between parametric and non-parametric bootstrap. Specifically, we follow this procedure: for $m = 1, \ldots, 10000$

(a) draw simulated values of professionalism and impartiality from independent normal distributions with mean equal to the reported value of the index for the country and standard deviation equal to the rough estimate (one fourth of the range reported for that country)

(b) for each country, sample with replacement four values from the vector with the simulated values of professionalism and impartiality and the observed values of statistical capacity and advanced degree holders

(c) calculate the index as the simple average of the resampled values

From the 10000 replications, we calculate and report the standard deviation and the empirical 2.5 and 97.5 percentile of the bootstrap distribution. Clearly, these are approximate estimates of uncertainty. In any case, strong inferences based on comparisons that are not close to statistical significance according to these approximate estimates of uncertainty should be avoided.[3]

(or simply puzzled by) the approach, hence we refrain from using it.

[3] Caution should also be exercised when dealing with the observations for which only the WB and UNESCO data is available. Given that (as explained in footnote 2) we do not imputate missing values, the standard errors estimated via bootstrap systematically underestimate uncertainity in the presence of missing values.

## 2.2 Efficacy

External know-how and capacity for innovation can be central to designing innovative and context-specific policies. One can capture the extent of such external sources of knowledge and expertise with data that measure think-tanks, top economics departments, policy schools, the number of researchers, and expenditure for research.

In order to assess the quality and vitality of the academic fields that mostly affect proposals for innovation and responses to new challenges in governance, we also identify data published by UNESCO on the full-time equivalent (FTE) number of researchers in a given country (both in general and specifically in the social sciences) per million inhabitants, and gross domestic expenditure for research (again, in general and for the social sciences) per million US dollars GDP. Finally, the (log) count of number of policy graduate programs in each country, as listed in the website GradSchools.com , serves as a rough measure of bureaucratic training.

**Think-tanks** To address the role of think-tanks, we identify three different sources. The first is the Global Go To Think Tanks Report compiled at the University of Pennsylvania. Information from the rankings (that reflect quality of the think-tanks in a given country) and the counts (that reflect quantity of think-tanks) is used. The second source is IDEAS RePEc, that reports a ranking of the top 25% of the think-tanks with members registered to the repository based on their scientific output. The third is NIRA, an organisation based in Japan that keeps an updated list of think-tanks around the world.

**Academic resources** Data on academic resources come from different sources. The ranking of economics departments at a global and regional level is the first source we use (IDEAS 2011). Specifically, we use two sets of rankings: the global rankings, and the various regional rankings. From each of the rankings, we calculate Borda-like scores as detailed below.

**A note of the creation of country scores from rankings** In this dashboard, we use both the ranking of think-tanks (in order of influence or quality) and of economics departments (in order of quality) to create country scores that aim at capturing the overall vitality of the research "industry" in a given

country. To achieve this goal, we use a mechanism similar to Borda scoring. Namely, we create organization-specific Borda scores from the ranking, and then aggregate them by country. As an example, consider the case of Canada. In a list of 50 think-tanks Canada has three think-tanks listed: the Fraser Institute, in position 16; the Centre for International Governance Innovation (CIGI) in position 32; the International Institute for Sustainable Development in position 48. Call $n_i$ the position of think-tank $i$ (from country $j(i)$) in a ranking of length $N$. We first assign to each think-tank the score $s_i = N + 1 - n_i$. Hence the Fraser institute receives a score of 35, CIGI a score of 19, and the International Institute for Sustainable Development a score of 3. The value of this variable for country $j^0$ is $x_{j^0} = \sum_{j(i)=j^0} s_i$. Hence the score for Canada is 35+19+3=57. From the scoring rule, it follows also that we attribute a score of one to the countries that have no think-tank listed in the ranking.

**Choice of weights** The aforementioned variables are aggregated (via weighted average of the centered and standardized scores) in an efficacy index. The Borda scores for economic departments and for think-tanks are logged before standardization and aggregation. In order to come up with a reasonable set of weights for the variables, we rely both on social-scientific and statistical intuition. First of all, we assign the same weight to each of two sub-indexes, one capturing think-tanks, one capturing the "official" academia and government-funded research. The 5 think-tanks variables are aggregated into the first sub-index, while the economics department information, the policy schools information, and the four variables from UNESCO belong to the second sub-index. Within each sub-index, we find it reasonable to downweight multiple measures from the same source. For this reason, equal weight is assigned to a) the mean of the three GO-TO Think Tanks-based scores b) the IDEAS-based score c) the NIRA adjusted count. For the academia sub-index, we assign equal weight to a) the score based on the IDEAS global ranking b) the score based on the IDEAS regional rankings c) the count of policy schools d) the four measures from UNESCO. This corresponds to weights reported in table 1.

As detailed in Stanig and Kayser (2013), there is no unique way of aggregating multiple variables in an index, and we encourage users to create their own indexes, to reflect both the purpose they have in mind and the conceptual framework they favour, using all or part of the variables we collected. The indexes we create have strong heuristic value but they should not be reified.

| Sub-index | Variable | Weight |
|---|---|---|
| Think-tanks | | .5 |
| | Score: UPenn, list of 75 | 1/18 |
| | Score: UPenn, list of 50 | 1/18 |
| | Count: UPenn, number of listed think-tanks | 1/18 |
| | Adjusted count: NIRA | 1/6 |
| | Score: IDEAS, list | 1/6 |
| Academia | | .5 |
| | Score: global econ dept ranking | 1/8 |
| | Score: regional econ dept ranking | 1/8 |
| | Policy schools | 1/8 |
| | UNESCO: researchers | 1/32 |
| | UNESCO: researchers in social sciences | 1/32 |
| | UNESCO: research funding | 1/32 |
| | UNESCO: research funding, social sciences | 1/32 |

Table 1: Weighting of the indicators in the efficacy index

**Estimates of uncertainty** For the standard errors, we rely on a bootstrap approach (Efron 1979; Efron and Tibshirani 1994). Given that none of the variables is treated as stochastic at the source, we treat the various indicators as repeated measures of the underlying phenomenon, and and we estimate the properties of the error term from the deviation of the values of the variables from the aggregate index. Formally, for each country $i$ we observe $K$ indicators $x_{ik}$ of the underlying phenomenon $\theta_i$. Each indicator can be thought of as an $x_{ik} = \theta_i + \epsilon_{ik}$ where $E(\epsilon) = 0$. We estimate $\theta_i$ simply by $\hat{\theta}_i = \frac{\sum_k w_k x_{ik}}{\sum_k w_k}$ where $w_k$ is the weight assigned to variable $x_k$. The uncertainty surrounding the estimate $\hat{\theta}_i$ depends on the covariance matrix of the $\epsilon$ terms. To avoid making too many assumptions about the distribution of the $\epsilon$, and following the approach of projects similar to ours (e.g., Transparency International until 2012, the Ibrahim Index of African Governance) we perform non-parametric bootstrap, country-by-country, and report the bootstrapped standard error (simply the standard deviation of the index taken over the 10000 bootstrap replications) and the empirical 2.5 and 97.5 percentiles of the bootstrap distribution. We also perform a Bayesian bootstrap (Rubin 1981). [4] We report the posterior standard deviation and the 95% highest posterior density intervals of the Bayesian bootstrap distribution.

[4] In substance, the results of the Bayesian bootstrap exercise are equivalent, in terms of implications, to the classical bootstrap.

## 2.3 Civil society

In order to estimate the strength of civil society, we rely on a mix of "objective" and self-reported (survey-based) measures.

From the Yearbook of International Organizations, we retrieve lists of all the organizations categorized in the Yearbook as "peace" organizations (this includes what, in more current terminology, is labeled as "human rights") and environmental organizations. We are able to learn, from the Yearbook, in which countries a given organization is present. We then count how many organizations of each type operate in a given country (and are not labeled as "inactive" by the YIO) . The scores reported in the dashboard are the raw counts, and they can be considered current as of late 2011/beginning of 2012.

For the participation, membership, and inequality measures, we rely on two survey collections that have sufficiently wide coverage and contain the items we need for the purpose, namely the WVS and the ISSP. We first create, for every respondent, summary measures of their self-reported membership and participation in civil society organizations. We adopt two definitions of civil society organizations, a narrow one (that includes only membership or activity in labor unions, political parties, environmental organizations, and charities) and a broad one (that includes also sport, art, professional, and "other" organizations). From these, we estimate country-level participation and membership rates, as the (survey-weighted) average of the individual scores (based on the "broad" definition).

To estimate inequality (or better, gradients) in participation and membership, we run regressions of the individual level participation and membership scores (based on the narrow definition) on a dummy for males, standardized placement in income stratification, and standardized education, country by country. Income and education are standardized by dividing by two standard deviations, so that they are approximately on the same scale as a dummy variable (Gelman 2008). Then, for each country, we sum the absolute values of the coefficients, to get an estimate of the overall strength of gender, income and education gradients on participation and membership in civil society. We treat these as our estimates of "inequality" in participation and membership. When a given country is covered by both the WVS and the ISSP, we take the average of the two scores for that country.

We then aggregate the six sources (the two counts, the participation and membership rates, and the

summaries of the gradients) by simple average, assigning equal weight to each source (and switching the sign of the inequality estimates, so that unequal participation is a sign of weakness of civil society).

To estimate standard errors, we are faced with the following situation: the YIO-based counts are clearly estimates of the number of organizations in the country, but they are constructed so that we only have one observation per country. On the other hand, the survey-based estimates have a standard error that (under relatively undemanding assumptions) can be used to assess the variability of the index itself.

To estimate the uncertainty around the point estimates of the civil society index, we follow this procedure. We first draw the counts from their posterior distribution under the non-informative priors suggested by Kerman (2011). In practical terms, the replication of a Poisson parameter $\lambda^{\text{rep}}$ is drawn from a $\text{Gamma}(x + \frac{1}{3}, 1)$ distribution where $x$ is the observed count. Then, we draw 1000 simulations $x^{\text{rep}} \sim \text{Poisson}(\lambda^{\text{rep}})$. For the survey estimates, the replications of both the proportions and the inequalities are drawn from independent normal distributions with mean equal to the survey estimate and standard deviation equal to the estimated standard error. Based on these 1000 replications of the inputs of the index, we then compute 1000 replications of the index. The standard deviation over the 1000 replications is reported as the standard error, and the 2.5 and 97.5 quantiles of the replications are reported as the bounds of the interval in the dashboard. While somewhat non-standard, this seems a reasonable approximation based on the approximate posterior distribution of the estimates with uninformative priors.

# 3 Transnational Governance Dashboard

The Transnational Governance Dashboard relies on several observable aspects of the transnational and international behaviour of countries.

## 3.1 UN treaties

In order to reconstruct different stances of countries we apply ideal point estimation to

the patterns of UN treaty ratification over the 1998-2012 period. The analysis exploits observed ratification patterns to infer underlying (and inherently unobservable) characteristics of countries. In the case in point, treaty ratification is used to estimate the latent willingness of countries to enter into international commitments.

The class of models estimated here starts from an observed matrix of zeros and ones (whether a country ratified a treaty or not, in this case) and extracts information about the latent predisposition of the country to ratify treaties. In the analysis of treaty ratification, ratification is considered as a "yea" vote on the provisions of the treaty, and a lack of ratification (or a refusal to sign the treaty in the first place) as a "nay" vote on its provisions.

The dimensionality of the space is inferred following an iterative process. First a uni-dimensional model is fitted. Identification (up to a 180-degree rotation) can be obtained by imposing the constraint that the ideal points have mean zero and standard deviation one. We then inspect the estimates and examine which treaties are not predicted well by the one-dimensional model, in the sense that the discrimination parameter is small in absolute value, and not statistically distinguishable from zero, and at the same time the treaty ratification is not lopsided. One of these treaties that is not well explained by the uni-dimensional model can be chosen as a good candidate to "anchor" a second dimension. This process can be repeated until the accuracy with which votes are explained is sufficiently high. As detailed in Stanig (2013), a two-dimensional model turns out to be appropriate for the treaty ratification data. In the dashboard we report the posterior mean and the posterior standard deviation of the ideal point of each country on each dimension. The models are estimated using the `ideal()` function in the `pscl` library (Jackman 2012 ) in the R environment.

## 3.2 Voting patterns in the United Nations General Assembly

The same type of models are also estimated on UN General Assembly (UNGA) votes on resolutions between December 2005 and December 2011. The UNGA roll calls up to the year 2009 come from the dataset maintained by Strezhnev and Voeten (2012-08), while for the most recent years, we collected and coded the information directly from the voting records posted by the United Nations on their

website (UNBISnet ). In the analysis, we treat abstention as equivalent to a 'nay' vote. As detailed in Stanig (2013), a four-dimensional model turns out to be appropriate, following the iterative procedure described above. We report, for each country, the posterior mean and the posterior standard deviation of the ideal point on each dimension.

### 3.3 International organisations for economic cooperation

The dashboard includes some measures that aim at evaluating the functioning of international organisations that deal primarily with economic cooperation.

**Voting power in the IMF Executive Board** The Transnational Governance Dashboard reports scores of voting power in the IMF Executive Board. A set of countries (US, France, and UK) are always included in the Executive Board, while others rotate, as representatives of groups. Unlike the United Nations General Assembly, the IMF Executive Board does not operate as a one-country, one-vote legislative body. Instead, each member country has a number of votes proportional to its economic importance. Depending on the rule adopted for a given decision, either a simple majority of voting weights or one of two types of supermajority is needed for a decision to be approved. The dashboard reports a score of voting power for each country in the Executive Board under each decision rule. The measures are based on the intuition that the power a country has depends on its ability to obtain concessions in exchange for support of a given decision. A country is defined as "critical" in a given winning coalition if the coalition would cease to achieve the required majority were that country to withdraw from the coalition itself. In other words, a country that is critical might be able to influence the other members of a potential majority by threatening to withdraw. We estimate the voting power of a given country as the ratio between the number of coalitions in which the country is included as a critical member, and the number of all possible minimal winning coalitions.

**WTO** In the Governance Report, we try to understand empirically what responsible sovereignty in trade policy –in the context of the institutions set up by the WTO/GATT– might entail. The Transnational Governance Dashboard reports some measures, based on averages over the 2006-2010 period

of data reported in Bown (2012a, 2012b. The first indicator looks at actions that countries take as victims of illegitimate trade practices of their trade partners. We count the number of Antidumping (AD) and Countervailing Duties (CVD) incidents listed in the Antidumping Database (Bown 2012a) and the Countervailing Duties Database (Bown 2012b). The value is the (log) number of antidumping measures that a country has taken against other countries. The figures we report are based on averages over the 2006-2010 period. The second indicator looks at the other side of trade behaviour, i.e. how often firms in a country are accused of engaging in dumping and related activities. The reported figures are also on log scale.

### 3.4 Global public goods

The concept of responsible sovereignty is closely related to contributions to the production of "global public goods". Data were collected on actual, observable contributions to the production of two global public goods: international peace and environmental protection.

**UN peacekeeping** Contributions to the UN peacekeeping missions are of two different sorts: monetary and in kind. Monetary contributions are composed of two parts, mandatory –assessed by the UN based on country GDP and size, with a special obligation by permanent members of the Security Council –and voluntary, on top of the assessed amount. The dashboard reports data both on monetary contributions and on the number of troops supplied by each country to peace operations over the years 2005-2011.

While it is not possible to collect data directly on military pay, we checked whether contributions can be predicted: by wages in manufacturing, as reported by the ILO; by a dummy for whether the country has compulsory military service or a volunteer-only army (own binary coding from the information available in the CIA Factbook); by military expenditures per soldier (estimated from the data about military expenditures in SIPRI and size of the military from the World Bank); and by population of the country. These variables, with the exception of country size, are not predictive of troop contributions. We adjust the number of troops contributed by each country over the 2005-2011 period only by population, with a regression of (log) troop contributions on (log) population. The

scores reported in the dashboard estimate the contributions that countries would be expected to make, all else equal, if they were all of the same size.

**Kyoto Protocol** In order to address a further manifestation of the willingness to contribute to the production of global public goods, and specifically the contribution to reduction of greenhouse gases, information was collected on the emission targets, and target fulfilment, for all the signatories of the Kyoto Protocol. Based on the data in we first calculate the change in emissions between respectively 2008, 2009, and 2010 and the base year. We then average these three changes to calculate the average change in emissions relative to the base year. In the dashboard we report the difference between this average and the target (that is also expressed in percentage reduction relative to the base year).

# 4 City Governance Dashboard

The third dashboard focuses on the governance of global cities. The dashboard relies mostly on surveys of entrepreneurs and ordinary citizens, complemented in selected cases with objective data collected by third-party organisations. WE isolate, in the World Bank Enterprise Surveys (World Bank n.d.), all the respondents located in a major city based on the information available in the surveys file. We identify respondents in 49 cities located in 48 countries. The enterprise surveys cover only countries outside of the group of advanced market economies with long-standing democratic regimes. WE also isolate respondents in global cities based on the information available in cross-national collections of surveys of citizens. Sources include many relatively recent surveys in the available cross-national collections: the International Social Survey Programme (ISSP) 2006 (ISSP Research Group 2008) and 2009 (ISSP Research Group 2012), the fourth wave of the World Values Survey (WVS 2009), the fourth round of the Afrobarometer (Afrobarometer 2010), wave two of the Asian Barometer, Latinobarometro 2009, and two Eurobarometer surveys (European Commission 2012a and 2012b). Respondents located in 73 cities in 64 countries were identified. Data from different sources are combined whenever possible (i.e. when comparable questions are asked in different survey collections) to achieve the broadest possible coverage. We estimate also the values for the rest of the country,

namely the averages for all respondents in the country that are not coded as dwellers of any of the major cities included in the analysis. The information coming from the enterprise and citizen surveys is combined with hard data compiled at the disaggregated subnational or city level and reported in the Mobility in Cities database (UITP 2006) and the OECD innovation statistics (OECD 2010). A detailed explanation of the conceptual framework of this dashboard is found in Stanig (2013). Here we focus only on the methodological aspects, while we refer the reader to the codebooks for variable definitions and sources.

## 4.1 City-level estimates and their standard errors

When we estimates levels of a phenomenon, we take the average of the individual answers to one or more survey items, in some cases after appropriate rescaling. Whenever survey weights are supplied with the survey data, we report weighted averages. For example, the level of generalized trust in a given city is estimated by the (weighted) proportion of respondents in that city that agree with a statement like "most people can be trusted". In some cases, we aggregate multiple survey items. For instance, to estimate perceived environmental quality, we compute the average of three responses to the World Values Survey, about air quality, water quality, and sanitation. The three survey items are combined in a scaled score (ranging in theory between 0 and 10) at the individual level. We then compute the (survey-weighted) average of the individual scores to come up with a city-level score. The standard errors reported in the dashboard are the square roots of the estimated sampling variance of the estimate. Whenever we estimate proportions, and these are such that the estimates are exactly equal to zero or to one (and therefore, the standard error would be zero), we add two "failures" and two "successes" to the observations when estimating the standard error.

## 4.2 Estimating "inequalities".

The city governance dashboard reports some quantities that, for the purpose of accessibility to the general policy-making public, we decided to dub "inequalities". Strictly speaking, these are the income gradients of certain interesting phenomena. For this purpose, linear regression models of individual

responses on position in the income distribution are estimated. In other words, we summarize how a given subjective perception (for instance, the belief that most people can be trusted) or an evaluation (how meritocratic a city is) is associated with the respondent's position in social stratification. The income variable is scaled so that it ranges from 0 to 1. We also report the standard errors of the regression coefficient.

### 4.3 Corruption index

From the surveys, several different measures of corruption at the city level are estimated. These reflect perceptions and victimisation, both of ordinary citizens and of entrepreneurs. From these, we also estimate a city-level corruption index, aggregating citizen and entrepreneur evaluations. One of the motivations for aggregation is to perform comparisons with a popular index of corruption perceptions, Transparency International's CPI, which is estimated at the country level and conflates petty and grand corruption in a single index.

In order to account for the fact that different sets of cities are evaluated by citizens and by entrepreneurs, we first impute the missing values on each variable using regression-based simple imputation). We then rescale each variable so that it has mean zero and standard deviation one half, including the imputed values in the computation of mean and standard deviation. After rescaling, we discard the imputed values. Imputing before rescaling is required: without such an adjustment, a score of zero on the entrepreneurs' evaluation (i.e. being the average city in the sample of –mostly developing– countries with Enterprise Survey data) is not equivalent to a score of zero on the citizens' evaluation (i.e. being the average city in the broader sample with citizen surveys, in which advanced countries are over-represented).

The reported standard error is simply the square root of the sum of the squared standard errors of the input variables.

### 4.4 Impartiality index

We have information about perceived impartiality in the entrepreneurs surveys and in the citizen surveys, and we combine these different sources in one single index by simple average of the standardized inputs. We first impute the missing values by regressing the entrepreneurs measures on the citizen measures, and vice versa. The simple-imputation values are used in the scaling step, when we center and standardize the input variables, but are discarded before estimation of the index.

The reported standard error is simply the square root of the sum of the squared standard errors of the input variables.

### 4.5 Different modes of answers in the Enterprise Surveys

In the Enterprise Surveys, some of the questions we use are answered, by different respondents, in one of two ways: either as a percentage of firm revenues, or as a level.[5] In order to aggegregate all available respondents, we convert the latter to percentage of firm revenues based on the firm revenues variable.

### 4.6 A tentative index of public transportation efficiency

To evaluate public transportation provision, we rely on the data collected by the Mobility in Cities database (UITP 2006), a data set aimed mainly at engineers and urban planners, which collects, for a selected number of cities, detailed information about public transportation provision and infrastructure investment related to transportation. In order to create a summary of the quality of public transit in a given city, we rely on four variables: the volume of public transit, its speed, its operating costs, and its energy consumption. The best public transit system is, ideally, the one that transports many people very quickly at low cost and at low energy consumption. In order to capture the fact that there are trade-offs and complementarities between each of these dimensions, we create an aggregate multiplicative index by computing the geometric mean of the inputs, appropriately rescaled. In practice, we

[5]For instance, a question reads " On average, what percent of total annual sales, or estimated total annual value, do establishments like this one pay in informal payments or gifts to public officials for this purpose?"

rescale the variables so that the lowest-scoring city receives a one, and the best-scoring city receives 100. We then average the (base 10) logs of the scores, and exponentiate the score, so that it "lives" again on a 1-100 scale. This is a very simple (and not widely adopted, to my knowledge) solution to an often-mentioned problem with indexes based on sums (or averages) for aggregation, in that they ignore complementarities. This index assigns a higher score to a city that performs relatively average on all dimensions than to a city that excels in half of the targets but performs poorly on the other half (for example, it has a very fast but small public transit system, and very energy-efficient but also very expensive system).

## References

Efron, Bradley. 1979. "Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics* 7:1-26.

Efron, Bradley, and Robert J. Tibshirani. 1994. *An Introduction to the Bootstrap* Chapman & Hall/CRC Monographs on Statistics Applied Probability.

European Commission. 2012a) *Eurobarometer* 72.2 (2009). TNS OPINION SOCIAL, Brussels [Producer]. GESIS Data Archive, Cologne. ZA4976 Data file Version 3.0.0, doi:10.4232/1.11137.

European Commission. 2012b. *Eurobarometer* 76.1 (2011). TNS OPINION SOCIAL, Brussels [Producer]. GESIS Data Archive, Cologne. ZA5565 Data file Version 2.0.0, doi:10.4232/1.11376.

Gelman, Andrew. 2008. "Scaling regression inputs by dividing by two standard deviations." *Statistics in Medicine* 27(15): 2865-2873.

Jackman, Simon. 2012. "`pscl`: Classes and Methods for R Developed in the Political Science Computational Laboratory." Technical document, Department of Political Science, Stanford University. Stanford, California. R package version 1.04.4. URL `http://pscl.stanford.edu/`

Kerman, Jouni. 2011. "Neutral Noninformative and Informative Conjugate Beta and Gamma Prior Distributions." *Electronic Journal of Statistics* 5:1450Ű1470.

Rubin, Donald B. 1981. "The Bayesian Bootstrap." *The Annals of Statistics* 9(1):130-134.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley.

Rubin, Donald B. 1996. "Multiple Imputation After 18+ Years." *Journal of the American Statistical Association* 91(434):473-489.

Stanig, Piero. 2013. "Governance Beyond the Nation-State: Estimating Governance Indexes at the Subnational and Transnational Level." In Helmut K. Anheier (Ed), *Governance Challenges and Innovations: Financial and Fiscal Governance.* Oxford: Oxford University Press.

Stanig, Piero, and Mark Kayser. 2013. " Governance Indicators: Some Proposals." In Helmut K. Anheier (Ed), *Governance Challenges and Innovations: Financial and Fiscal Governance.* Oxford: Oxford University Press.