New results on the influence of climate on the distribution of population and economic activity

Füssel, Hans-Martin
Potsdam Institute for Climate Impact Research

04. March 2009

New results on the influence of climate on the distribution of population and economic activity

Hans-Martin Füssel *(fuessel@pik-potsdam.de)*

Potsdam Institute for Climate Impact Research

March 2009

Abstract

This paper applies G-Econ+, an updated version of the G-Econ database by Nordhaus, to analyze the influence of climatic and geographic factors on the geographic distribution of population and economic activity. I discuss options for improved treatment of several statistical problems associated with G-Econ, which are not addressed adequately in the original G-Econ analysis. Reanalysis of key results from the original G-Econ analysis corrects some surprising results therein. Extensive sensitivity analysis determines the robustness of the relationship between climatic factors and economic activity across alternative central estimators. Further analysis assesses revealed climatic preferences of population, the effects of climate parameters on different quantiles of economic variables, and synergies between temperature and precipitation. I find that population density has a much stronger influence on output density than output per capita. Furthermore, least developed countries are located in a climatic zone where all indicators of economic activity decline with increasing temperature.

Keywords:

Climate; macroeconomics; population; cross-sectional analysis; G-Econ

JEL Codes:

C21; Q54; C82

Copyright Notice:

1 Introduction

The economic prosperity of regions is determined by a complex interplay of geographic, climatic, environmental, historical, political, institutional, and cultural factors. While the importance of climate for human welfare has long been recognized [Mills 1942, Diamond 1999, Sachs 2003], most research in the last decades has attempted to explain the wide divergence in wealth by cultural and political factors, such as colonial legacy [Landes 1998, Acemoglu et al. 2001, Engerman and Sokoloff 2005, Rodrik et al. 2004]. Recognition of the severity of anthropogenic climate change, however, has lead to a renaissance of research on the influence of climate on economic conditions[1]. In particular, several recent studies have extrapolated the observed relationship between climatic and economic factors to estimate the impacts of climate change on global and regional economic productivity and wealth [Nordhaus 2006, Nordhaus 2008, Dell et al. 2008, Dell et al. 2009].

G-Econ 1.3 combines a variety of data sources to estimate key climatic, geographic, demographic and economic variables for all 1°-by-1° terrestrial grid cells of the world. [Nordhaus 2006] (abbreviated as PNAS in the remainder of this paper). This database combines (i) climate data, which are available for each grid cell based on direct observations and interpolations in data-poor regions [New et al. 2002]; (ii) population data, which are available for most grid cells from a variety of sources, including census data and extrapolations from night-time lights [Balk and Yetman 2004]; (iii) data on economic output in local currency, which have generally been collected at the first subnational level for most large countries; and (iv) currency conversion factors to the US$ based on market exchange rates and purchasing power parities at the national level.

[Nordhaus 2006, Nordhaus 2008] applies the G-Econ database (i) to describe the influence of individual climatic and geographic variables on economic output *per area*, (ii) to describe the influence of individual climatic variables on economic output *per capita*, and (iii) to estimate the impact of climate change on global and regional economic output based on a cross-sectional analysis similar to the Ricardian technique for estimating economic impacts of climate change in agriculture[2] [MENDELSOHN et al. 1994]. Independent studies have analyzed the impact of climate variability and change on income distribution and economic growth based on panel data at the national level [Dell et al. 2008] and on cross-sectional data at the national, state, and municipal level [Dell et al. 2009].

This paper has been motivated by some counter-intuitive results in PNAS and by a more general interest in the implications of several statistical challenges associated with the G-Econ database. Section 2 identifies the main statistical problems with G-Econ, reviews their consideration in PNAS, and discusses various methods to address them more appropriately. Section 3 analyzes the influence of individual climatic and geographic factors on economic output *per area* and *per capita*. Most results presented there reanalyze and extend results from PNAS related to the above-mentioned applications (i) and (ii); reanalysis of the climate impact estimates from PNAS –application (iii)– was not possible due to insufficient documentation of the multivariate regression equations applied. Section 4 presents additional statistical analysis based on G-Econ+, which sheds further light on the (often complex) relationship between climatic factors and the distribution of population and economic activity.

[1] In this paper, I use the term “influence” to describe any statistically significant relationship between explanatory and explained variables. It does not necessarily imply a direct causal relationship between these variables.

[2] *“Using the G-Econ database, we can estimate the impact of different warming scenarios on output using our gridded global database. […] The projection of the impact of climate change relies upon an equation with the natural logarithm of gross cell output density as a dependent variable and geophysical variables as independent variables. Additionally, I have added variables that are country-specific linear temperature effects.”* [Nordhaus 2008]

Section 5 summarizes the main findings, discusses their implications for the robustness of results in PNAS, and concludes with suggestions for further research.

2 Statistical problems

G-Econ presents data on several explanatory variables (alternatively known as regressor, predictor or dependent variable; here: climatic and geographic factors) and explained variables (alternatively known as regressands or predictand or independent variable; here: economic variables) for all 1°-by-1° terrestrial grid cells of the earth. Data on the explained variables (specifically: output per area and output per capita) is associated with several statistical problems, which have important implications for descriptive statistics and regression analysis.

2.1 Strong positive skew and heteroskedasticity

Output density (i.e., economic output per area) and output per capita are strongly right-skewed and they exhibit strong heteroskedasticity (i.e., the variance of these variables depends on the explanatory variable). Strong skewness has important implications for descriptive statistics because alternative measures of central tendency can differ significantly. Even the question which of two climatic regimes is associated with higher output density may depend on the choice of central estimator (see Section 3 for examples). Therefore, the comparison of aggregated features of different subpopulations from a strongly skewed dataset (e.g., output density in different climate regimes) should not be based on a single central estimator. Skewness and heteroskedasticity bring about various challenges for regression analysis, such as making insignificant variables appear to be statistically significant. They can be addressed by transforming the raw data (e.g., with a Box-Cox power transformation, of which the log transformation is a special case) with retransformations allowing for heteroskedasticity [DUAN 1983, Johnson et al. 1994, Manning 1998]; by using robust regressions such as quantile regression [Koenker 2005]; by using alternative weighting approaches based on exponential conditional models and generalized linear models [Manning and Mullahy 2001]; and by approximating the data with non-normal probability distributions such as the generalized gamma distribution; [Manning et al. 2005]. PNAS does not mention skewness and heteroskedasticity explicitly but applies a modified log transformation to all explained variables, which was apparently introduced to address these problems (see Section 2.5 for a discussion).

2.2 Excess zeros

G-Econ data on output density exhibit “excess zeros” (i.e., there are many grid cells with zero population and economic output). Excess zeros cause several problems for statistical analysis, including biased estimates of the effects of explanatory variables and overestimation of the dispersion of the explained variable. Furthermore, excess zeros may complicate the application of data transformations to address heteroskedasticity. In particular, the log transformation cannot be applied directly because the logarithm is not defined for zero values of the explained variable. Excess zeros can be addressed by censoring data (i.e., excluding zero observations); by two-part models, which are generally distinguished into conditional models (also known as hurdle models) [Welsh et al. 1996] and mixture models (also known as zero-inflation models) [LAMBERT 1992]; and by other methods such as modified two-part models [Mullahy 1998]. PNAS deals with zero values in an explained variable by increasing them to a small non-zero value before the log transformation. The problems associated with this cutoff are discussed in Section 2.5.

2.3 Different natural weights of data points

The data points in G-Econ represent different land area because the size of a 1°-by-1° grid cells decreases with absolute latitude, coastal grid cells contain less land area, and grid cells with national borders are split into several data points. Therefore, averaging of output density across grid cells requires the application of area weights. Averaging of output per capita may apply population or area weights, depending on the purpose of the analysis (see Sections 3.4 and 4.4). Population-weighting of grid cells corresponds to picking a person randomly from a given climate zone and estimating their output per capita whereas area-weighting corresponds to picking a unit area randomly from a given climate zone and estimating the output per capita of its inhabitants. The difference between weighted and unweighted central estimators is generally small for output density but it can be substantial for output per capita. PNAS applies equal weights when aggregating explained variables across grid cells.

2.4 Different spatial resolution of explanatory and explained variables

The explanatory and explained variables in G-Econ are measured at different spatial resolutions. Climatic, geographic, and demographic variables are available at the level of 17,491 grid cells whereas output per capita is only available at the level of 4,095 administrative units, some of which comprise hundreds of grid cells. Estimates of the explained variable in data-poor regions can be substantially biased if the explained variable varies systematically within administrative units as a function of the explanatory variable. In G-Econ, this problem is most relevant for the estimation of output per capita in very cold regions. The importance of this problem can be assessed by analyzing the sensitivity of results to different spatial aggregation units and to consider the resulting differences in the interpretation of results (see Section 3.4). G-Econ+ enables such a sensitivity analysis by providing all data at the level of grid cells and administrative units (see Section 3.2). PNAS does not consider the different spatial resolution of explanatory and explained variables in the interpretation of results.

2.5 Log-transformation of explained variables

The statistical analysis in PNAS applies the log-transformed explained variables rather than the raw data, apparently to reduce skewness and heteroskedasticity[3]. This log transformation, however, causes several new problems. Most importantly, it can lead to systematically biased regression estimates for the raw data if the heteroskedasticity is not accounted for in the retransformation of the dependent variables. In the words of [Manning 1998] (p. 285), *"there is a very real danger that the log scale results may provide a very misleading, incomplete, and biased estimate of the impact of covariates on the untransformed scale, which is usually the scale of ultimate interest"*. Second, the log transformation is undefined for grid cells with zero output density. PNAS applies a cut-off value of 1 US$/km^2 to output density before the log transformation; this cutoff value is arbitrary and can have substantial influence on the results (see Sections 3.1and 3.3). Third, the log-transformation with cutoff introduces substantial scale-dependency into the aggregation of output density (and to a lesser degree, of output per capita)[4]. Hence, the regression of log-transformed output density to biophysical

[3] In the case of output per capita, an alternative motivation for the log-transformation could have been that the logarithm of output per capita is often used as a measure of the instantaneous utility of income or consumption. These isoelastic utility functions are particularly widespread in integrated assessment models of climate change (see [DeCanio 2003], Table 2.4). This motivation does not hold for the logarithm of output density, which does not correspond to any known social welfare metric.

[4] For example, assume four neighbouring grid cells of 100 km^2 each, one of which produces an annual output of 40 Mio US$, whereas the other three are unpopulated. The average log-transformed output density with cutoff of these four grid cells is ¼*log10(40,000,000/100)+3/4*log10(1)=1.4. If these four grid cells are merged into one

conditions may be highly sensitive to the size of the analysis units. None of these problems is discussed in PNAS.

2.6 Summary

G-Econ data are associated with a variety of statistical problems, which have important implications for descriptive statistics and regression analysis. Each of these problems can be addressed by statistical methods that are already applied in various contexts (e.g., in health economics). Addressing all problems together, however, poses substantial problems for statistical analysis and involves difficult trade-offs. Therefore, the choice of statistical methods needs to be guided by the purpose of the statistical analysis and by the underlying hypothesis governing the phenomenon of interest. PNAS does not discuss the implications of any of these problems for statistical analysis and the robustness of the results presented. The method applied to address two of them (skewness and heteroskedasticity) introduces new problems, which are not discussed either. For that reason, it remains an open question whether it is possible to reliably estimate future economic impacts of global climate change based on the extrapolation of regression coefficients derived from a multivariate linear regression of cross-sectional data on climate, geography, and economic activity.

3 Climatic and geographic influence on economic activity

In this section I analyze the influence of individual climatic and geographic factors on output density and output per capita. Most of the results presented here reanalyze or directly extend results from PNAS. Section 3.1 shows that some surprising results in PNAS have been caused by flawed analysis techniques and reveals the large sensitivity of some results to alternative central estimators. Section 3.2 introduces G-Econ+, an updated database that corrects several errors in G-Econ 1.3 and that is available at two different spatial resolutions. The following subsections present additional analyses of the influence of temperature on output density (Section 3.3) and on output per capita (Section 3.4).

3.1 Climatic and geographic determinants of output density

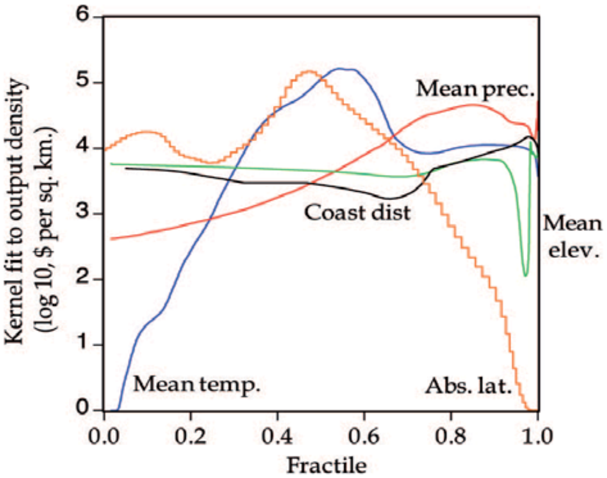


Figure 1: Original quantile plots for key geographic variables and output density (Source: [**Nordhaus 2006**], Fig. 2; used with permission; © 2006 PNAS).

Figure 1 depicts the relationship between several geographic and climatic variables and output per area as determined in PNAS. The explanatory variables are normalized by their quantiles

larger grid cell, however, the log-transformed output density of this merged grid cells is 4/4*log10(40,000,000/400)=5.0.

and the explained variable is log-transformed, presumably to reduce its skewness and heteroskedasticity. Since the logarithm of zero is undefined, grid cells with zero output are included as having an output density of 1 US$/km^2. The underlying dataset comprises Greenland but not Antarctica because climate data for the latter is not available. Most of the curves agree with common knowledge about the geographic distribution of population and economic activity. The suggested maximum of economic activity furthest away from the coasts and the shape of the elevation curve, however, are counter-intuitive.

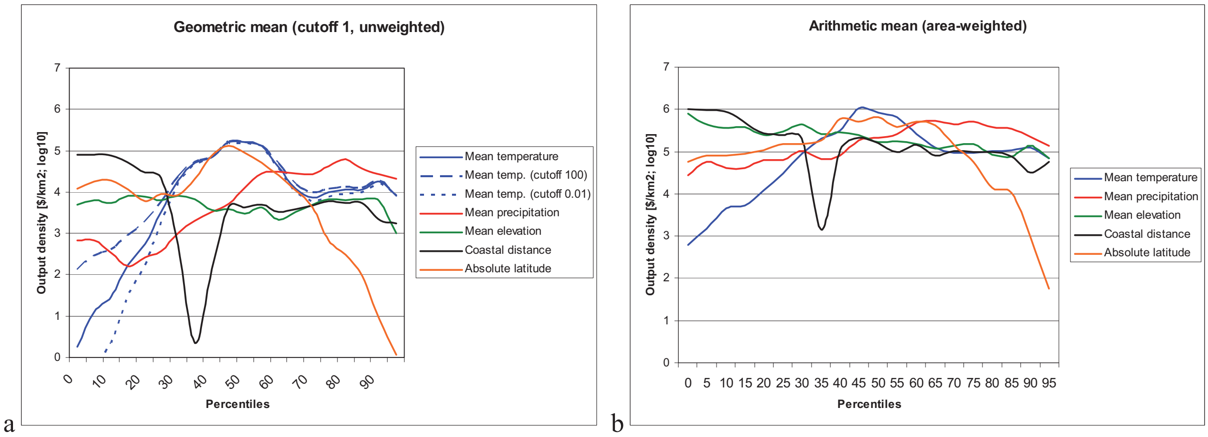


Figure 2: Reanalyzed quantile plots for key geographic variables and output density, applying two different methods for aggregating output density across grid cells (see text).

Figure 2.a depicts the results of a reanalysis based on the same data and methods as in PNAS. For compatibility with the original analysis, I also applied the same arbitrary truncation for unpopulated grid cells. The main difference between this diagram and the one in Figure 1 concerns the quantile plots for coastal distance, which show no agreement at all. The reanalysis curve is much more plausible than the original one[5], which suggests an unknown error in the PNAS analysis[6]. The differences in the curves for precipitation and elevation may be due to different smoothing methods applied in the two analyses. The curves for temperature and latitude are in perfect agreement. The dashed blue lines show that the shape of the temperature plots is clearly sensitive to variations in the cutoff value, in particular in very cold regions.

Determining average output density for quantiles of the explanatory variables by taking the unweighted mean of the truncated log-transformed data is problematic for several reasons. First, it causes a systematic bias in estimates of the non-transformed variable due to its heteroskedasticity (see Section 2). Second, the results are highly sensitive to the cut-off value (see Figure 2.a) and to the spatial aggregation level (not shown here). Third, grid cells with small land area (i.e., those at high latitudes, in coastal regions and in border regions) are overrepresented in the sample. Figure 2.b shows the results of an alternative analysis where average output density was determined as the area-weighted mean of the non-transformed data. The quantile plots for non-transformed output density exhibit several important differences compared to those for the log-transformed variable (depicted in Figure 2.a). First of all, maximum output density for all explanatory variables is about one order of magnitude larger. Second, the variation in output density across the explanatory variables is generally smaller than in the original analysis (except for elevation). The variation becomes similar, however, if a larger cutoff value of 100 US$/km^2 is applied to the log-transformed data (see

[5] The sharp drop in output density around the 40th percentile of coastal distance is caused by a clustering of unpopulated grid cells from Greenland around this percentile.

[6] I have presented the initial findings of this reanalysis to W. Nordhaus but his reply was unable to clarify the reasons for the differences in results.

the example for mean temperature in Figure 2.a). Third, the fractile plot for elevation shows a clear downward trend. Further differences can be detected for the temperature and precipitation plots.

None of the quantile plots for the log-transformed data that have been applied in the regression analysis in PNAS shows a clear monotonous effect on the explained variables. PNAS has considered the non-monotonous influence of temperature on output density by using temperature and temperature squared as predictors in the regression equation but all other climatic and geographic variables were applied only in their linear form. The plots for the non-transformed data show a decreasing trend for elevation and for coastal distance (if Greenland is excluded from the analysis), a mostly increasing trend for precipitation, and non-monotonous trends for temperature and absolute latitude.

3.2 Development of G-Econ+

The analysis above has relied on the original G-Econ 1.3 database in order to separate the effects of alternative analysis methods from the effects of changes in the underlying database. A detailed review of G-Econ has identified a number of data errors in G-Econ 1.3 [Füssel 2008][7]. For that reason, I have developed an updated version of G-Econ 1.3, denoted as G-Econ+. The main improvements over G-Econ 1.3 are the application of consistent currency exchange rates in China and the USA (they vary by several orders of magnitude in G-Econ 1.3), the elimination of inconsistencies between different variables in the same grid cell (e.g., when a grid cell has zero population but non-zero output), of inconsistencies between variables for different grid cells in G-Econ and of inconsistencies between G-Econ and the underlying country files. G-Econ+ is available at two spatial resolutions: for all 17,491 land-based grid cells and for the 4,095 administrative units distinguished in G-Econ 1.3. For a detailed description of G-Econ+, see [Füssel 2008].

3.3 Climate influence on output density

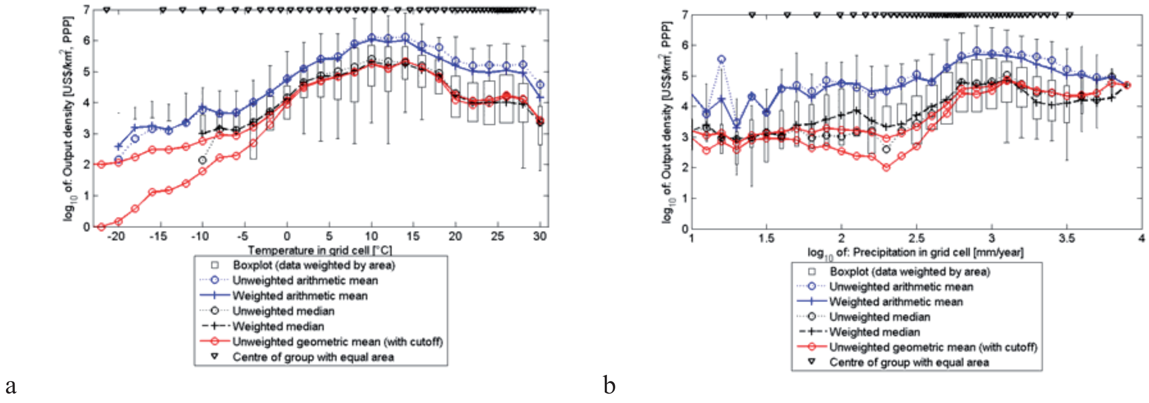


Figure 3: Influence of temperature (a) and precipitation (b) on output density (see text for details).

Figure 3 depicts the influence of temperature and precipitation on several central estimators of output density. The boxes depict the 25th and 75th percentile of output density (weighted by land area) within each temperature bin, and the whiskers depict the 5th and 95th percentiles. The geometric mean (red curves) was calculated for two different cutoff values (1 and 100 US$/km^2). The black triangles at the top depict the mean temperature of 50 regions with equal land area (see Section 3.4 for a detailed discussion).

[7] After completion of G-Econ+, Nordhaus published an updated version of G-Econ (G-Econ 2.11, http://gecon.yale.edu/documents/GEcon_211_121608_post.xls), which may have addressed some of the data problems described here. The technical paper describing G-Econ+ is included as an Appendix to this paper.

Figure 3.a closely resembles Fig. 4 in PNAS where the unweighted geometric mean (lower red curve) was applied as the central estimator but the additional quantiles and central estimators provide a wealth of additional information. The solid red and blue curves are similar to the "mean temperature" curves in Figure 2.a and b except for the use of temperature bins rather than quantiles. Output density assumes a maximum in the 10-14°C temperature bins with slight variations across central estimators. Most estimators show another local maximum for the 26-28°C bins (see below for further discussion). The arithmetic mean is much larger (by about one order of magnitude) than the modified geometric mean across most of the temperature domain. For a suitable choice of the cutoff value (about 100 US$/km^2), the modified geometric mean is similar to the median except for very low temperatures (below -10°C) where the median becomes zero. The variation in output density within temperature bins is large; most interquartile ranges span more than one order of magnitude.

According to Figure 3.b, output density shows no clear trend up to about 200 mm annual precipitation (see below for further discussion), increases up to about 800-1200 mm and decreases for even higher precipitation. Most central estimators show minimum output density for very low precipitation. The arithmetic mean is much larger (by an order of magnitude or more) than the modified geometric mean across most of the precipitation domain. For a suitable choice of the cutoff value (about 100 US$/km^2), the modified geometric mean is similar to the median. The variation in output density within temperature bins is about as large as for temperature; most interquartile ranges span more than an order of magnitude. The variation in output density across precipitation regimes is less pronounced than across temperature regimes. Furthermore, the median of output density never becomes zero. These findings suggest that unfavourable precipitation is a less stringent constraint on the population of a region than unsuitable (i.e., very cold) temperature.

Area-weighting of grid cells has little impact on the results in both diagrams except for very warm and very wet regions where the unweighted arithmetic mean of output density is somewhat larger than the weighted mean. Most likely, the former gives too much weight to densely populated coastal grid cells whose land area is generally smaller than that of inland grid cells.

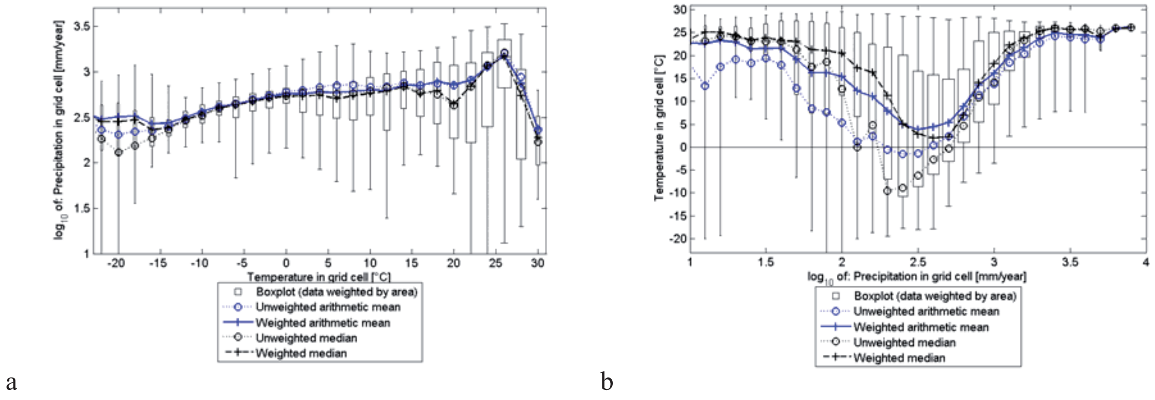


Figure 4: Relationship between temperature and precipitation (both ways).

Several central estimators in Figure 3 exhibit a local maximum of output density around 26°C and a local minimum around 200 mm precipitation per year. What is the reason for these regional extrema? To shed further light on this question, Figure 4 explores the covariance between temperature and precipitation. Figure 4.a shows that mean precipitation increases slightly with temperature up to a maximum around 26°C, and decreases substantially for even warmer temperature. This result suggests that the local maximum in output density around 26°C is primarily caused by very high precipitation rather than by the particularly favourable temperature (see Section 4.4 for further analysis). Figure 4.b shows a U-shaped relationship between precipitation and temperature. The driest and the wettest regions are much warmer than average, whereas medium-low precipitation occurs mostly in temperate and cold regions. Very cold regions are most strongly represented around 250 mm precipitation per year. This result suggests that medium-low precipitation is not less favourable for economic activity than very low precipitation *per se* but that medium-low precipitation is more often associated with very cold temperatures that are strongly unfavourable for economic activity.

3.4 Climate influence on output per capita

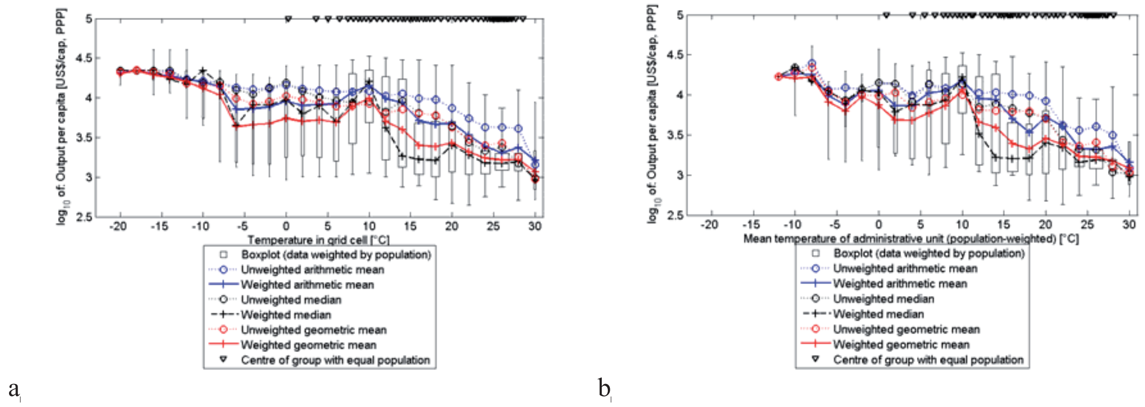


Figure 5: Influence of temperature on output per capita for different spatial aggregation units of temperature data (a: grid cells; b: administrative units).

Figure 5 depicts the influence of temperature on output per capita for different spatial aggregation units. Figure 5.a shows both variables at their original resolution (i.e., grid cells for temperature and administrative units for output per capita). The black triangles on top depict the mean temperature of 50 regions with equal population. The location of the left-most triangle close to 0°C means that the mean temperature in the homeland of the "coolest" 2% of global population is about 0°C. Regions outside the range spanned by the first and last triangle cover only a small fraction (less than 1%) of global population. For that reason, considerable caution should be applied when interpreting the relationship between temperature and economic activity outside this range.

Figure 5.a closely resembles Fig. 3 in PNAS where the unweighted geometric mean (red curve) was applied as the central estimator. All estimators suggest that output per capita assumes a maximum in the coldest regions (below -10°C). Population-weighting has a large effect on estimates of output per capita for a given temperature. Population-weighted estimates are generally much lower than unweighted and area-weighted estimates (not shown here), except for the sparsely populated regions below -6°C. Furthermore, all weighted estimators assume a local maximum at moderate temperatures (around +10°C), which is not present for the unweighted and area-weighted estimators. In other words, population living around 10°C is wealthier on average than population in any other climate zone above -8°C even though population in an average 0°C grid cell is wealthier than population in an average 10°C grid cell. This "population-area-paradox" suggests that the 0°C temperature bin contains more sparsely populated regions with high output per capita and fewer densely populated regions with low output per capita than the 10°C bin. The 10°C temperature bin is exceptional because the weighted median exceeds the unweighted median and the (weighted and unweighted) arithmetic mean, and the 25th percentile is much higher than for any other temperature bin above -8°C. These findings indicate that densely populated regions are on average wealthier than sparsely populated regions and income inequality across regions is comparatively small around 10°C. The variation of output per capita within temperature bins is still substantial but somewhat smaller than for output density; the same holds for the difference between arithmetic and geometric mean.

A possible interpretation of the substantial difference between the strongly fluctuating population-weighted and the smoother unweighted (and area-weighted) curves involves distinguishing two different population groups (or economic sectors). Output per capita of the first group is strongly dependent on climate whereas that of the second is not. The area-weighted median is dominated by the first group, which is spread rather evenly within a given climatic zone; the population-weighted median is more strongly affected by the second group, which is more heterogeneously distributed and concentrated in regions that provide a climate-insensitive income basis. Income of the second group tends to be lower than that of the first group except around 10°C. Quantitative examination of this hypothesis would require additional data on the relative importance of different economic sectors and is beyond the scope of this paper.

The finding that output per capita is highest in very cold regions is somewhat surprising. PNAS discusses this finding but does not provide a definite explanation. One possible explanation involves the strong prevalence of capital-intensive activities in very cold regions, such as mining and oil extraction, but this hypothesis cannot be examined further because data on GDP from oil extraction in the G-Econ database is incomplete and often inconsistent [Füssel 2008]. The finding may also be interpreted in a reverse manner in the sense that most of the population living in very cold regions requires high output per capita to cope with (or compensate for) the unfavourable climate. The result might also be an artefact of the difference in spatial resolution between the explanatory and explained variables (see Section 3.2). This hypothesis is examined in Figure 5.b where the spatial resolution of climatic and economic data was made consistent by assigning grid cells to temperature bins based on the population-weighted average temperature of the administrative unit they belong to. Figure 5.b and Figure 5.a are very similar for the data-rich temperature bins above 2°C. In addition, output per capita still assumes a maximum in the coldest temperature bins (below -6°C) for all estimators. In Figure 5.b, however, output per capita according to the unweighted estimators is about as high or higher in the 10°C temperature bin than in colder grid cells (except for the very sparsely populated region below -6°C). Furthermore, the very cold temperature bins below -12°C that exhibit maximum output per capita in Figure 5.a are no longer represented in Figure 5.b.[8]

In summary, regions below -6°C are so sparsely populated that it does not appear to be justified to analyze the relationship between temperature and output per capita below this temperature level. The relationship between temperature and output per capita between -6°C and 10°C varies across central estimators, weighting schemes, and spatial aggregation units. The majority of estimators show a rather flat relationship with a (local) maximum at 10°C. All estimators agree that output per capita tends to decline above 10°C.

The influence of precipitation on output per capita (not shown here) is weaker than that of temperature. Output per capita shows no clear trend up to about 150 mm annual precipitation, increases up to about 300-800 mm and decreases for even higher precipitation. Most central estimators show similarly low output capita for very low and very high precipitation.

3.5 Climate-output reversal

PNAS notes *"opposite relationships between climate and output depending on whether we look at output per person or output per area"* (p. 4). This "striking paradox" is labelled the

[8] The coldest populated administrative subunit in G-Econ and G-Econ+ is Tunu on Greenland with an area-weighted temperature of about -21°C. The population-weighted temperature of Tunu, however, is -3°C, because the population of Tunu is centred in the comparatively mild coastal regions. The coldest region represented in Figure 5.b is the Nunavut territory in Canada with a mean population-weighted temperature of about –12°C.

"climate-output reversal"[9], and considerable attempt is made at explaining it. What can the present reanalysis contribute to this discussion? First of all, even in the PNAS analysis the climate-output reversal does not apply above 10°C because both output indicators decrease substantially (by about one order of magnitude) above this temperature level. This reanalysis confirms the finding that output density increases with temperature up to 10°C but it disagrees that there is a robust relationship between temperature and output per capita up to 10°C. PNAS finds a negative relationship between temperature and output per capita, based on one unweighted central estimator that relates variables at different spatial aggregation units. The reanalysis in Section 3.4 finds a highly variable relationship between temperature and population-weighted output per capita (at the same spatial aggregation level) between -6°C and 10°C with a positive trend for population-weighted estimators and without a clear trend for unweighted and area-weighted estimators. The specific finding in PNAS that output per capita is highest in the coldest regions and decreases smoothly with increasing temperatures is not robust across alternative statistical methods.

In a nutshell, the influence of temperature on output density and its components can be characterized as follows. Below 10°C, there is no clear trend in output per capita whereas population density increases strongly and monotonously with rising temperature (see Section 4.3). As a result, output density also increases strongly and monotonously. Above 10°C, output per capita decreases monotonously with rising temperature; population density also has a negative trend but exhibits a pronounced local maximum around 26-28°C. As a result, output density shows a negative trend with a small local maximum around 26-28°C.

4 Additional statistical analysis based on G-Econ+

Most of the results presented in Section 3 reanalyze or directly extend results presented in PNAS. G-Econ+ can be applied in many more ways to analyze the influence of climatic factors on the distribution of population and economic activity. In this section I assess the revealed climatic preferences of population (Section 4.1), the effects of climate parameters on different quantiles of economic variables (Section 4.2), the relative importance of the two determinants of economic output density (Section 4.3), and the combined influence of temperature and precipitation on the distribution of population and economic activity (Section 4.4).

4.1 Climatic preferences at the regional level

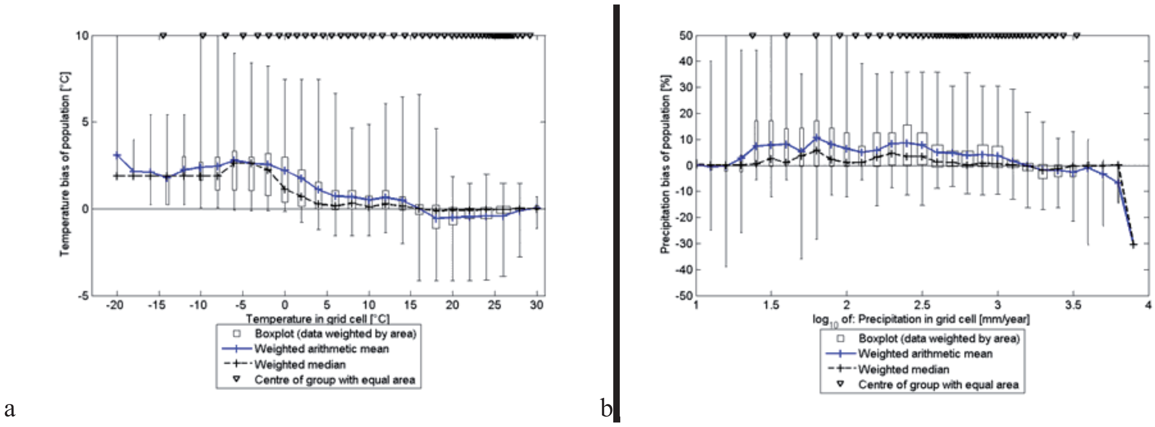


Figure 6: Climatic preferences of population within administrative units. The plots depict the difference between population-weighted and area-weighted temperature (a) and precipitation (b) across the whole range of temperature and precipitation, respectively.

[9] I retain this term here even though "temperature-output reversal" would be a more accurate denotation of the proposed phenomenon.

The availability of G-Econ+ at two different spatial resolutions enables analysis of the influence of climatic factors on population distribution at a regional level. Figure 6 depicts the climatic preferences of population within administrative units, whereby the horizontal axis refers to the actual climate within a grid cell and the vertical axis to the difference between population-weighted and area-weighted temperature and precipitation.[10] Figure 6.a shows that the temperature bias of population is positive up to 16°C and slightly negative above that level. In other words, population prefers to live in the warmer parts of an administrative unit up to an annual mean temperature of 16°C and in the cooler parts above that temperature level. Figure 6.b shows that the precipitation bias of population is positive up to about 1500 mm/year and slightly negative above that level. In other words, population prefers to live in the wetter parts of an administrative unit up to a precipitation of 1500 mm and in the drier parts above that precipitation level. The magnitude of the climate bias for different temperature and precipitation regimes should be interpreted with much caution as it depends on factors such as the size and climatic heterogeneity of administrative units in G-Econ+, both of which vary systematically with the explanatory variable. It would be interesting to perform a similar analysis using economic weights rather than population weights but this would require economic data at a higher resolution than in G-Econ.

4.2 Climate influence on quantiles of economic activity

The influence of temperature and precipitation on output density and output per capita was already investigated in Sections 3.3 and 3.4. The temperature influence on output density is rather robust across central estimators, as shown by the similar shape of the respective curves in Figure 3.a. The temperature influence on output per capita, in contrast, varies substantially across central estimators (see Figure 5.a). In this section I present quantile plots that provide additional information on the influence of temperature on economic activity.

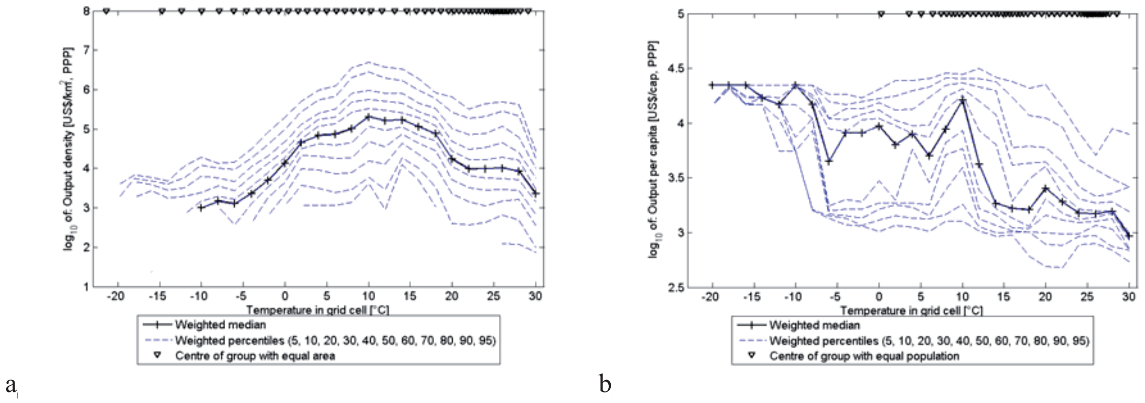


Figure 7: Influence of temperature on various quantiles of output density (a) and output per capita (b).

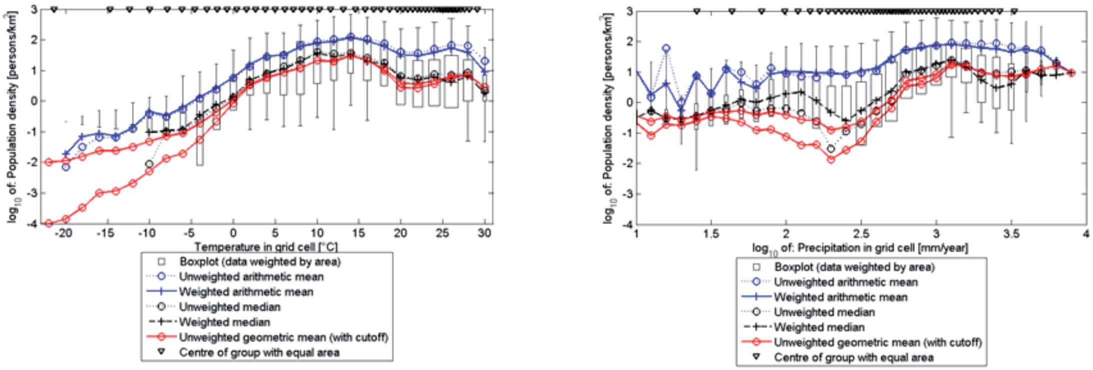
Figure 7 depicts the influence of temperature on various quantiles of output density and output per capita. The "weighted median" curves in Figure 7.a and Figure 7.b correspond to the respective curves in Figure 3.a and Figure 5.a, respectively. According to Figure 7.a, temperature has a similar effect on all quantiles of output density. The lower quantiles are undefined for very cold temperature (median: below -10°C; 10th percentile: below 0°C), which indicates the large share of unpopulated grid cells in these climatically unfavourable regions. According to Figure 7.b, the influence of temperature on output per capita is much weaker than on output density. All quantiles are defined because unpopulated grid cells are

[10] Comparison of population-weighted and area-weighted climate eliminates regression toward the mean that would occur if the population-weighted climate of an administrative unit was compared with the actual climate of the containing grid cells.

excluded from the analysis. Low quantiles (up to the 10th percentile) and high quantiles (80th percentile and higher) are rather insensitive to a change in temperature within the range from -6°C to 14°C even though the median varies considerably. The lower quantiles are most sensitive to very cold temperature, where output per capita is very high. Hence, most of the population inhabiting very cold regions (below -8°C) lives in wealthy administrative units (for further discussion see Section 3.4). In contrast, the higher quantiles are most sensitive to warm temperature, where output per capita is relatively low. Hence, the warmer the temperature (above a level of 10°C), the smaller is the share of population living in wealthy administrative units.

4.3 Output density versus population density

Output density is the product of population density and output per capita. How important is the variation in these two factors for the distribution of output density across regions and climatic regimes? The rank correlation of output density with population density and output per capita across all grid cells with non-zero population is 0.890 and 0.079, respectively. Hence, the distribution of economic output is dominated by the distribution of population; output per capita plays a significant but clearly secondary role.



a

b

Figure 8: Influence of temperature (a) and precipitation (b) on population density.

Figure 8 depicts the influence of temperature and precipitation on population density. The curves of the central estimators resemble those of Figure 3, which depicts the influence of temperature and precipitation on output density. Comparison with Figure 3 reveals that temperature and precipitation have a similar effect on population density and output density except that the decrease for above-optimal temperature and precipitation is less pronounced, and the local maximum in the 26-28°C temperature bins is more pronounced for population density. The strong relationship between output density and population density implies that knowledge on the climatic, geographic, historical and other determinants of population distribution can go a long way in explaining the distribution of economic activity; it is unlikely that the determinants of economic activity differ substantially from those of population distribution.

4.4 Combined influence of temperature and precipitation on the distribution of population and economic activity

The analysis up to now has focused on the influence of individual geographic and climatic factors on the distribution of population and economic activity. These factors, however, do not act in isolation. As a complement to the preceding analysis, the analysis in this section focuses on the combined effects of temperature and precipitation.

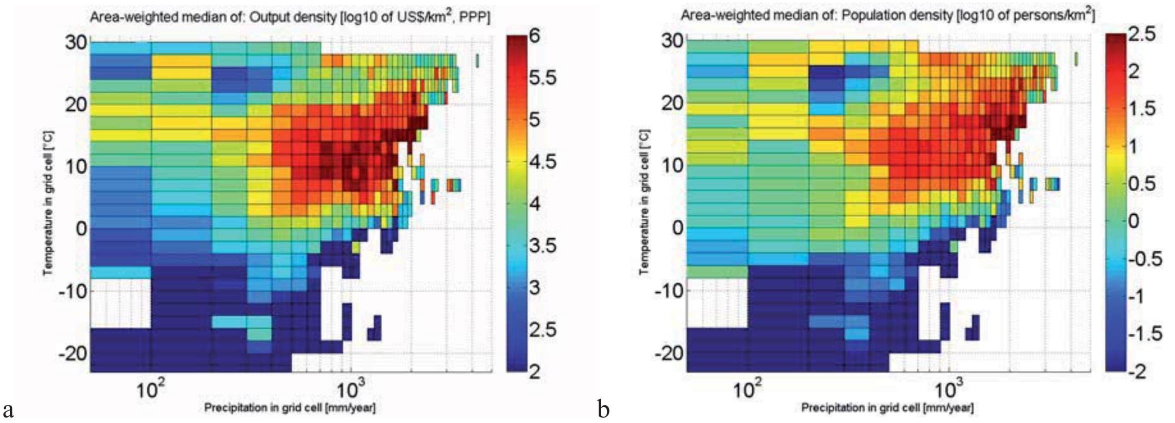


Figure 9: Combined influence of temperature and precipitation on output density (a) and population density (b).

Figure 9 shows the distribution of the area-weighted median of output density and population density across the full range of temperature and precipitation. Other central estimators (not shown here) exhibit a similar pattern. According to Figure 9.a, output density is largest in areas with precipitation above 500 mm/year and temperature between 6 and 20°C (somewhat warmer for large precipitation); it is smallest below -6°C. More precipitation tends to increase output density across the full temperature range; the relationship between temperature and output density is inverse U-shaped for most precipitation levels. The pattern of population density (one of the two determinants of output density) in Figure 9.b is very similar to the pattern of output density in Figure 9.a, strengthening the findings from Section 4.3.

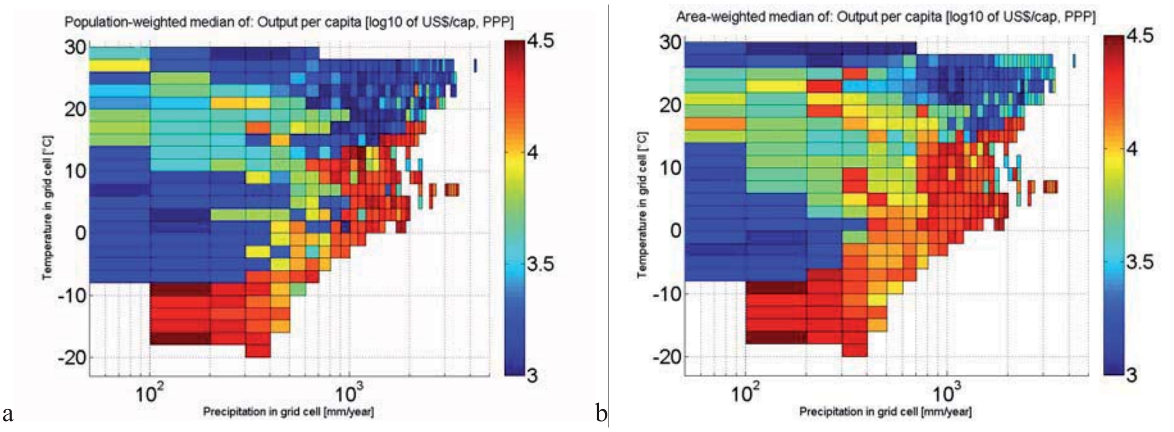


Figure 10: Combined influence of temperature and precipitation on output per capita weighted by population (a) and by area (b).

Figure 10 depicts the distribution of output per capita across the full range of temperature and precipitation; Figure 10.a applies population weights and Figure 10.b area weights in the calculation of the median. Figure 10.a shows two distinct climate regimes with very high output per capita: first, the very cold and sparsely populated regions below -8°C (see the discussion in Section 3.4); and second, temperate and moist regions between 0 and 18°C with precipitation above a temperature-dependent threshold level. There are also two distinct climate regimes with very low output per capita: first, cool and dry regions with temperature between -8 and 10°C and precipitation below 300 mm/year; and second, warm and moist regions with temperature above 18°C and precipitation above a temperature-dependent threshold level. More precipitation tends to increase output per capita up to 16°C and to decrease it above 18°C. Increasing temperature tends to decrease output per capita for precipitation above 500 mm/year and has a varied effect in drier regions. In dry regions with precipitation below 300 mm/year, warm temperature (above 10°C) is associated with higher output per capita than cool temperature (between -8 and 10°C), which is in contrast to the negative relationship between temperature and output per capita at higher precipitation levels. Further analysis (not shown here) reveals that the warm and dry regions comprise many coastal areas whereas the cool and dry regions are predominantly located inland. Hence, the positive temperature effect on output per capita in dry regions appears to be strongly influenced by confounding non-climatic factors.

Figure 10.b applies area-weighting rather than population-weighting in the aggregation of output per capita across grid cells with similar climate conditions. The results are similar to Figure 10.a, but there are two important differences. First, the pattern of the area-weighted mean is generally "smoother", showing fewer abrupt changes between neighbouring climate zones than the population-weighted median, which is more sensitive to the effect of large population centres. Second, the area-weighted median is generally larger than the population-weighted median. This effect is particularly strong in cool-temperate regions with a temperature between -6 and 10°C and precipitation between 300 and 800 mm/year. In this climate regime, the area-weighted median is determined by rather wealthy regions in Canada whereas the population-weighted median is determined by less wealthy regions in Russia.

5 Summary and conclusions

This paper has discussed the main statistical challenges for analysis with G-Econ (Section 2), reanalyzed key findings from the original G-Econ analysis (Section 3), and provided additional information on the complex relationship between biophysical factors and the distribution of population and economic activity (Section 4). In this section I summary the main methodological and empirical results and provide some suggestions for further work.

5.1 Summary of methodological findings

This paper identified a number of challenges for statistical analyses of the relationship between climatic factors and the distribution of economic activity based on G-Econ, including strong skewness and heteroskedasticity of the explained variables, excess zeros in data on output density, variations in area and population across grid cells, and different spatial resolutions of the underlying datasets. Each of these statistical problems has been addressed individually in the scientific literature, in particular in health economics, but addressing them together still presents substantial challenges for statistical analysis. It is therefore important to tailor the statistical analysis to the underlying question of interest, and to assess the robustness of results across alternative analysis methods.

The arithmetic mean, quantiles (such as the median), and the geometric mean without cutoff (for datasets without excess zeros) are all internally consistent central estimators of explained variables in G-Econ; their suitability depends on the specific goal of the statistical analysis. The arithmetic mean is appropriate to investigate which climate zones are most strongly represented in global economic output and population but it is highly sensitive to confounding non-climatic factors. Quantiles are generally more robust to the influence of non-climatic factors but they may exhibit discontinuities when applied to data-sets with excess zeros (such as for population density and output density). Application of the geometric mean *without* cut-off to output per capita can be interesting because it corresponds to the expected logarithmic utility of consumption, a utility function that is often used in integrated assessment models of climate change. The geometric mean *with* cut-off, as applied to output density in PNAS, is highly sensitive to the cutoff value and to the size of the spatial unit; it should therefore not be used as the basis for regression analysis.

Equal weighting of all grid cells (as in PNAS) does not appear defensible because of the large differences in area and population across grid cells. Area-weighting should be applied for determining the effect of biophysical factors on population density and output density. The calculation of central estimates for output per capita may apply either population or area weights, depending on the question of interest. Area-weighted output per capita is less sensitive to the effects of population centres where large fractions of the population may be engaged in climate-insensitive economic activities but it over-represents people living in sparsely populated regions.

Grid cells are the appropriate unit of analysis for the relationship between climate and population density. They are also the most appropriate unit of analysis for the relationship between climate and output density, because the variation in output density is more strongly determined by the variation in population density than in output per capita. Analysis of the relationship between climate and output per capita at the level of grid cells may be systematically biased, in particular for climate regimes with low population numbers (such as very cold regions). Such analysis should therefore be complemented by an analysis conducted at the level of administrative units.

If zeros are neglected, the log transformation applied in PNAS can reasonably well make data on population density and output density normal distribution-like. Regression of these variables, however, has to account for excess zeros and correct the effects of the log transformation on central estimates. The log transformation and other Box-Cox transformations are less successful in making data on output per capita normal distribution-like, even though these data do not contain zeros.

5.2 Summary of empirical results

The main empirical results of this reanalysis are as follows. Output density and population density assume a maximum in the 10-14°C temperature range and for precipitation around 1000 mm/year, with slight variations across central estimators and weighting schemes. Very cold temperatures are a stronger constraint on output density and population density than very warm temperatures or insufficient precipitation. All estimators of population density and most estimators of output density exhibit a local maximum in the 26-28°C temperature range, which coincides with maximum precipitation. The arithmetic mean of output density for a given temperature range is about one order of magnitude larger than the geometric mean with cutoff as applied in PNAS but the shape of the relationship with temperature is not substantially altered. Output density and population density exhibit a maximum in the 800-1200 mm/year precipitation range. Regression of output density on climatic factors is complicated by the presence of excess zeros; its explanatory power is hampered by the large variation of output density within temperature and precipitation bins and across central estimators. The relationship between precipitation and output density is more sensitive to the choice of central estimator than that for temperature and output density. Counterintuitive results in PNAS suggesting that output density assumes a maximum furthest away from the coasts have been corrected. Regression analysis needs to consider that most biophysical variables in the G-Econ database have a non-monotonous effect on output density.

Most estimators of output per capita are fairly insensitive to variations in temperature below 10°C; they show a (local or global) maximum at 10°C and a decreasing trend for higher temperatures. The exact location of global and local maxima varies across estimators. The effect of temperature is strongest for intermediate quantiles of output per capita (e.g., the median of grid cells within a given temperature range) whereas high and low quantiles are less sensitive. Estimates of output per capita for very cold regions appear unreliable due to a lack of representative economic data; they are subject to additional uncertainty due to the incomplete coverage of output from oil and gas extraction in G-Econ, which is the dominant economic activity in many of these regions. The combined effect of temperature and precipitation on output per capita is complex, which can be partly explained by large covariance between these two climate variables as well as between climate and other biophysical variables (e.g., coastal distance).

Output density is the product of population density and output per capita. Climatic factors have a much stronger influence on population density than on output per capita. As a result, the geographical and climatic distribution of output density closely resembles that of population density.

5.3 Robustness of results in Nordhaus (2006)

This reanalysis provides a more accurate, but also considerably more complex, description of the influence of climatic and geographic factors on the distribution of economic productivity than the original analysis in PNAS even though it relies on the same primary data. The reanalysis of key results from PNAS presented here involved correcting erroneous data in the G-Econ database, updating flawed analyses, and analyzing the sensitivity of key results to alternative methods for data aggregation. The PNAS results on the relationship between temperature and economic output could largely be confirmed but the breadth of information presented here allows for much better assessment of the robustness of this relationship. The "climate-output reversal" postulated in PNAS is not a robust result of this reanalysis; it is contingent on the particular choice of central estimators in PNAS.

PNAS has estimated the economic impacts of global warming by extrapolating into the future a multivariate regression of log-transformed output density on climatic and geographic factors from a cross-sectional analysis. The data errors in G-Econ 1.3, the various statistical problems not addressed in PNAS, the sensitivity of the climate-output relationship to the choice of central estimator, and the lack of monotonicity in the relationship between most explanatory and explained variables in G-Econ raise serious doubts regarding the robustness of this climate impact estimate.

5.4 Concluding remarks

Climatic and geographic factors are important determinants of the geographic distribution of population and economic activity. Hence, the interest in using information about the current distribution of these variables to draw conclusions for the implications of global climate change is understandable. Simply extrapolating a (not necessarily robust) relationship identified in a particular cross-sectional analysis into the future without any explanatory model, however, does not appear to be the best way forward. The analysis presented here provides a wealth of results on the relationship between climatic factors and the distribution of economic activity, which helps to distinguish robust relationships from correlations that are contingent on a particular method of statistical analysis. One robust result is that all indicators of economic activity decline with an increase in annual mean temperature beyond 14°C. This result is consistent with climate-response functions developed for key climate-sensitive sectors in the USA based on experimental as well as cross-sectional evidence [Mendelsohn and Schlesinger 1999]. Incidentally, this climatic regime comprises all least developed countries with the exception of parts of Afghanistan, Nepal, and Bhutan. Identification of "non-robust" relationships is also valuable as it can guide the choice of suitable statistical methods (e.g., quantile regression) and the specification of more appropriate causal models (e.g., distinguishing climate-sensitive from climate-insensitive economic activities).

6 Acknowledgements

I am grateful to Bill Nordhaus for making the G-Econ data publicly available and for his willingness to openly discuss my questions at the 3rd Atlantic Workshop on Energy and Environmental Economics (A Toxa, Spain, 4-5 July 2008) where I presented initial results of the analysis described here. I thank Michael Hanemann, Carlo Carraro and Georg Feulner for discussions on a previous version of this paper. Andreas Neumann, Alexander Haumann and Thomas Cornelissen helped with data input and analysis.

7 References

[Acemoglu et al. 2001] D Acemoglu, S Johnson, and JA Robinson. The colonial origins of comparative development: An empirical investigation. *AMERICAN ECONOMIC REVIEW*, 91(5):1369–1401, DEC 2001. ISSN 0002-8282.

[Balk and Yetman 2004] D. Balk and G. Yetman. The Global Distribution of Population: Evaluating the Gains in Resolution Refinement. Documentation for GPW Version 3, 2004. URL http://beta.sedac.ciesin.columbia.edu/gpw/docs/gpw3_documentation_final.pdf.

[DeCanio 2003] S. J. DeCanio. *Economic Models of Climate Change*. New York, NY, 2003.

[Dell et al. 2008] Melissa Dell, Benjamin F. Jones, and Benjamin A. Olken. Climate Change and Economic Growth: Evidence from the Last Half Century. Working Paper 14132, National Bureau of Economic Research, June 2008. URL http://www.nber.org/papers/w14132.

[Dell et al. 2009] Melissa Dell, Benjamin F. Jones, and Benjamin A. Olken. Temperature and Income: Reconciling New Cross-Sectional and Panel Estimates. Working Paper 14680, National Bureau of Economic Research, January 2009. URL http://www.nber.org/papers/w14680.

[Diamond 1999] J. Diamond. *Guns, Germs, and Steel: The Fates of Human Societies*. W. W. Norton, New York, 1999.

[DUAN 1983] N DUAN. SMEARING ESTIMATE - A NONPARAMETRIC RETRANSFORMATION METHOD. *JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION*, 78(383):605–610, 1983. ISSN 0162-1459.

[Engerman and Sokoloff 2005] Stanley L. Engerman and Kenneth L. Sokoloff. Colonialism, Inequality, and Long-Run Paths of Development. Working Paper 11057, National Bureau of Economic Research, January 2005. URL http://www.nber.org/papers/w11057.

[Füssel 2008] Hans-Martin Füssel. The G-Econ+ database on climate and economic activity: data and methods, 2008. URL http://www.pik-potsdam.de/ fuessel/gecon/gecon_plus_description.pdf.

[Johnson et al. 1994] N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*, volume 1. Wiley, New York, second edition, 1994.

[Koenker 2005] Roger Koenker. *Quantile Regression*. Cambridge Universtiy Press, New York, 2005.

[LAMBERT 1992] D LAMBERT. ZERO-INFLATED POISSON REGRESSION, WITH AN APPLICATION TO DEFECTS IN MANUFACTURING. *TECHNOMETRICS*, 34(1):1–14, FEB 1992. ISSN 0040-1706.

[Landes 1998] D. Landes. *The Wealth and Poverty of Nations*. W. W. Norton, New York, 1998.

[Manning 1998] WG Manning. The logged dependent variable, heteroscedasticity, and the retransformation problem. *JOURNAL OF HEALTH ECONOMICS*, 17(3):283–295, JUN 1998. ISSN 0167-6296.

[Manning and Mullahy 2001] WG Manning and J Mullahy. Estimating log models: to transform or not to transform? *JOURNAL OF HEALTH ECONOMICS*, 20(4):461–494, JUL 2001. ISSN 0167-6296.

[Manning et al. 2005] WG Manning, A Basu, and J Mullahy. Generalized modeling approaches to risk adjustment of skewed outcomes data. *JOURNAL OF HEALTH ECONOMICS*, 24(3):465–488, MAY 2005. ISSN 0167-6296.

[Mendelsohn and Schlesinger 1999] R Mendelsohn and ME Schlesinger. Climate-response functions. *AMBIO*, 28(4):362–366, JUN 1999. ISSN 0044-7447.

[MENDELSOHN et al. 1994] R MENDELSOHN, WD NORDHAUS, and D SHAW. THE IMPACT OF GLOBAL WARMING ON AGRICULTURE - A RICARDIAN

ANALYSIS. *AMERICAN ECONOMIC REVIEW*, 84(4):753–771, SEP 1994. ISSN 0002-8282.

[Mills 1942] Clarence A. Mills. *Climate Makes the Man*. Harper Brothers, New York, 1942.

[Mullahy 1998] John Mullahy. Much ado about two: reconsidering retransformation and the two-part model in health econometrics. *Journal of Health Economics*, 17(3):247 – 281, 1998. ISSN 0167-6296. doi: DOI: 10.1016/S0167-6296(98)00030-7. URL http://www.sciencedirect.com/science/article/B6V8K-41CX5B8-1/2/741fdd96c438bad1b8f1f18eff9cdbae.

[New et al. 2002] M New, D Lister, M Hulme, and I Makin. A high-resolution data set of surface climate over global land areas. *CLIMATE RESEARCH*, 21(1):1–25, MAY 23 2002. ISSN 0936-577X.

[Nordhaus 2006] WD Nordhaus. Geography and macroeconomics: New data and new findings. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 103(10):3510–3517, MAR 7 2006. ISSN 0027-8424. doi: 10.1073/pnas.0509842103.

[Nordhaus 2008] William D. Nordhaus. New Metrics for Environmental Economics: Gridded Economic Data. *The Integrated Assessment Journal*, 8:73–84, 2008.

[Rodrik et al. 2004] D Rodrik, A Subramanian, and F Trebbi. Institutions rule: The primacy of institutions over geography and integration in economic development. *JOURNAL OF ECONOMIC GROWTH*, 9(2):131–165, JUN 2004. ISSN 1381-4338.

[Sachs 2003] Jeffrey D. Sachs. Institutions Don't Rule: Direct Effects of Geography on Per Capita Income. Working Paper 9490, National Bureau of Economic Research, February 2003. URL http://www.nber.org/papers/w9490.

[Welsh et al. 1996] AH Welsh, RB Cunningham, CF Donnelly, and DB Lindenmayer. Modelling the abundance of rare species: Statistical models for counts with extra zeros. *ECOLOGICAL MODELLING*, 88(1-3):297–308, JUL 1996. ISSN 0304-3800.

Appendix

The G-Econ+ database on climate and economic activity: data and methods

Hans-Martin Füssel *(fuessel@pik-potsdam.de)*

Potsdam Institute for Climate Impact Research
October 2008

Abstract

G-Econ 1.3 (Nordhaus, 2006)[11] is a gridded database of climatic, geographic, demographic, and economic variables. This paper presents a modified version of G-Econ, denoted as G-Econ+, which corrects several errors in G-Econ 1.3[12]. G-Econ+ is available in a gridded version and at the level of (national and subnational) administrative units.

1. Introduction

G-Econ 1.3 provides data on climatic and geographic factors, population, and economic output[13] for all land-based 1°x1° grid cells. Climate data in G-Econ are for the 1961-1990 period, population and economic data refer to 1990 but were rescaled to the administrative boundaries of 2000 where appropriate. Data on climate, geography, and population was available on a gridded basis whereas economic data (i.e., output per capita) was only available at the level of administrative units (except for Canada). G-Econ+ provides data on land area, climate (annual mean temperature and precipitation), population, and gross product (market exchange rate and purchase power parity) at two spatial resolutions: for all 17,491 land-based 1°x1° grid cells with climate data in the CRU_CL_2.0 dataset; and for all 4,095 national and subnational administrative units[14] included in G-Econ 1.3. The main goal for the development of G-Econ+ was to produce a database that is internally consistent at both spatial resolutions. Results of an analysis based on G-Econ+ have already been presented at the 3rd Atlantic Workshop on Energy and Environmental Economics[15] but this paper focuses on the development of G-Econ+.

G-Econ+ combines data from G-Econ 1.3 and from the individual country files, both of which are kindly made available at the G-Econ homepage (http://gecon.yale.edu/). G-Econ provides data at the level of "grid cells by country"[16] but it does not contain information on the

[11] W. D. Nordhaus (2006): Geography and macroeconomics: New data and new findings. Proceedings of the National Academy of Sciences 10.1073

[12] After completion of G-Econ+ and this technical paper, Nordhaus published an updated version of G-Econ (G-Econ 2.11, http://gecon.yale.edu/documents/GEcon_211_121608_post.xls), which may have addressed some of the data problems described here.

[13] Analogous to Nordhaus, the terms "economic output" and "gross product" are used interchangeably here.

[14] Note that the level of disaggregation varies widely across and within the 189 countries considered. For instance, more than a quarter of the subnational administrative units belong to three countries only: Nigeria (538 subunits), Germany (438 subunits), and Ghana (141 subunits).

[15] A Toxa, Spain, 4-5 July 2008, http://webs.uvigo.es/rede/toxa/pages/3rd-atlantic-workshop/program-papers.php

[16] The expression "grid cell by country" means that G-Econ contains one entry for each combination of grid cell (characterized by latitude and longitude) and country. Most grid cells belong to only one country and are represented by a single entry in G-Econ. Data on land area, population, and gross product of multi-national grid cells, however, are provided in separate entries for each national fraction of that grid cell. The term "grid cell by subnational unit" is used in an analogous way.

subnational unit(s) that a grid cell belongs to. The country files provide data at the level of "grid cells by subnational unit" but they are only available for 92 out of 190 countries covered by G-Econ. G-Econ+ combines data from G-Econ and the country files for two reasons: first, to be able to provide all data at the level of subnational administrative units; and second, to identify and correct apparent errors in G-Econ. The merging of these data sources was complicated by differences in the variables included in G-Econ and the country files, by various data gaps and inconsistencies, and by differences in the file structure across country files.

2. Description of G-Econ+

Tables 2 and 3 describe the variables contained in the two versions of G-Econ+:

Variable	Units	Explanation
LAT	°	Latitude of SW corner of grid cell
LONG	°	Longitude of SW corner of grid cell
AREA_cell	km^2	Land area of grid cell
POP_cell	persons	Population of grid cell
TEMP_cell	°C	Mean temperature of grid cell
TEMP_av_area	°C	Mean temperature of administrative unit (area-weighted)
TEMP_av_pop	°C	Mean temperature of administrative unit (population-weighted)
PREC_cell	mm/month	Mean precipitation of grid cell
PREC_av_area	mm/month	Mean precipitation of administrative unit (area-weighted)
PREC_av_pop	mm/month	Mean precipitation of administrative unit (population-weighted)
GCPMER_cell	US$/year	Gross cell product (market exchange rate)
GCPMER_AREA	US$/(km^2*year)	Output density (market exchange rate)
GCPMER_POP	US$/(person*year)	Output per capita (market exchange rate)
GCPPPP_cell	US$/year	Gross cell product (purchasing power parity)
GCPPPP_AREA	US$/(km^2*year)	Output density (purchasing power parity)
GCPPPP_POP	US$/(person*year)	Output per capita (purchasing power parity)
POP_density	persons/km^2	Population density

Table 2: Variables contained in the grid cell version of G-Econ+

Variable	Units	Explanation
COUNTRY_ID	—	Country ID (arbitrary)
SUBUNIT_ID	—	Administrative unit ID (arbitrary)
AREA_subunit	km^2	Land area of administrative unit
POP_subunit	persons	Population of administrative unit
Dummy_1	—	
TEMP_av_area	°C	Mean temperature of administrative unit (area-weighted)
TEMP_av_pop	°C	Mean temperature of administrative unit (population-weighted)
Dummy_2	—	
PREC_av_area	mm/month	Mean precipitation of administrative unit (area-weighted)
PREC_av_pop	mm/month	Mean precipitation of administrative unit (population-weighted)
GCPMER_subunit	US$/year	Gross product of administrative unit (market exchange rate)
GCPMER_AREA	US$/(km^2*year)	Output density (market exchange rate)
GCPMER_POP	US$/(person*year)	Output per capita (market exchange rate)
GCPPPP_subunit	US$/year	Gross product of administrative unit (purchasing power parity)
GCPPPP_AREA	US$/(km^2*year)	Output density (purchasing power parity)
GCPPPP_POP	US$/(person*year)	Output per capita (purchasing power parity)
POP_density	persons/km^2	Population density
ADMIN	—	Name of first-level administrative subunit
DISTRICT	—	Name of second-level administrative subunit
SUBDISTRICT	—	Name of third-level administrative subunit

Table 3: Variables contained in the administrative unit version of G-Econ+

3. Development of G-Econ+

The main steps in the development of G-Econ+ were as follows:

Land area

Data on land area in G-Econ generally appears reliable but some grid cells have zero land area even though economic output is non-zero. Data on land area at the level of grid cell by country was taken from G-Econ, whereby cells with zero area were excluded from further analysis. Within each grid cell, land area was allocated to administrative subunits[17] according to the "Rate in grid" (RIG) data from the from country files.

Population

Population at the level of grid cell by country was taken from G-Econ, which is very similar to the GPW v3 dataset (http://sedac.ciesin.columbia.edu/gpw/) but excludes very sparsely populated regions. Population data is handled inconsistently in the country files. In most country files, population at the level of grid cell by country corresponds to the *sum* of all entries for subnational units. In 16 country files, however, it corresponds to the *average* of all (identical) entries for subnational units. If consistent population data at the level of grid cell by subunit was not available in the country files, G-Econ+ allocates grid cell population to different subunits according to their share in economic output or in land area, depending on data availability.

Climate

G-Econ+ uses climate data (mean temperature and precipitation) from G-Econ, which appears identical to the CRU_CL_2.0 dataset (http://www.cru.uea.ac.uk/cru/data/tmc.htm). There are some inconsistencies between G-Econ and the country files, mostly in regions with steep climate gradients, which may be related to different spatial interpolation methods. Climate data is missing for all of Antarctica and for a few other grid cells, which were excluded from further analysis.

Gross product

G-Econ+ generally uses economic output data in US$ at the level of grid cell by country from G-Econ. For some countries, however, data on gross product in US$ in G-Econ contain very substantial errors caused by the application of erroneous currency exchange rates (see Section 4 for details). Therefore, gross product data for USA, China, and Angola were taken from the country files. Gross product at the level of grid cell by subunit is available in some country files but not in others. For most European countries, for instance, gross product in US$ at the level of grid cell by country is reported in a single arbitrary subunit whereby all other subunits contain zero entries. If economic data in local currency is available for each subunit and the exchange rate (i.e. the ratio of gross product in US$ and in local currency, summed up to the level of grid cells) is identical across all grid cells within a country, gross cell product in US$ was allocated to subunits within a grid cell according to the distribution of gross product in local currency. For countries with inconsistent exchange rates, gross product in US$ at the level of grid cell by subunit was determined by allocating gross cell product from G-Econ to different subunits according to their share in population or in land area, depending on data availability.

[17] The term subunit is used here to denote the lowest-level administrative unit for which data on GDP per capita is available in G-Econ 1.3. It comprises nations as well as ~4,000 first-level (for most major countries), second-level (for some countries), and third-level (only for Sudan) subnational administrative units.

G-Econ+ reports the same output per capita in all grid cells within an administrative subunit. To this end, the procedure above needs to be modified for Canada, where data on output per capita was collected at the level of grid cells rather than administrative subunits and in many grid cells with land-based or ocean-based oil extraction. In these cases, average output per capita in each administrative subunit excluding oil extraction[18] was calculated as follows. In those grid cells where GCPOILMER (i.e., GCP from oil extraction, based on market exchange rates) and GCPNOMER (i.e., GCP from other activities) sum up to GCPMER (i.e., total GCP), GCPNOMER was used. If GCPOILMER is available but inconsistent with GCPMER or if GCPMER exceeds 500.000 US$ per capita[19], the grid cell was excluded from the calculation of average output per capita. In addition, GDP per capita for Alaska was set to the average value of the USA.[20] An analogous procedure was applied for gross product based on purchase power parities (PPP) rather than market exchange rates (MER).

4. Data problems in G-Econ

Any effort to produce a global database comprising climatic, environmental, demographic and economic data faces a multitude of challenges regarding data availability and quality. This fact is clearly acknowledged in the documentation of the G-Econ database. Some inconsistencies in the G-Econ database, however, appear to be related to flaws in data aggregation rather than to limitations of the primary data. The extensive consistency checks performed during the development of G-Econ+ revealed four types of data problems:

1. Inconsistencies between variables in the same grid cell (e.g., a grid cell has non-zero output but zero population).
2. Inconsistencies between variables for different grid cells of the same database (e.g., currency exchange rates in G-Econ differ between grid cells of the same country).
3. Inconsistencies between identical variables in G-Econ and the country files (e.g., grid cell population differs substantially between G-Econ and the country files); and
4. Other data problems (e.g., population numbers that appear unrealistically high or low even though they are consistent between G-Econ and the country files).

This section starts with a description of one important data problem related to inconsistent exchange rates. The remainder of this section mentions other data problems briefly. Note that for the sake of brevity not all variable names are explicitly explained in the latter part of this section.

Inconsistent currency exchange rates

Country	min(GCPMER / GCPLC)	max(GCPMER / GCPLC)	Deviation occurs also in grid cells without oil extraction?
China	0; 0.03744	8636	Yes

[18] The decision not to include output from oil extraction in G-Econ+ was based on two reasons: first, oil extraction often occurs in areas without (permanent) population, leaving output per capita (i.e., per inhabitant) undefined; secondly, current oil extraction appears insensitive to global climate change, which is the main motivation for the development of G-Econ and G-Econ+.

[19] This threshold is only exceeded for grid cells where the country files mention economic output from oil extraction but where this information has not been transferred to G-Econ.

[20] GDP per capita in G-Econ is far higher in Alaska than in any other US state, largely due to oil extraction, but oil extraction is not mentioned separately in the G-Econ database.

USA	0.3357	14.96	Yes
Kuwait	2.471	1802	Yes
Yemen	0.0074	0.0356	Yes
Syria	0.02182	0.08534	Yes
Saudi Arabia	0.2758	0.3005	Yes
Iran	0.00164	0.00472	No
Libya	3.412	3.899	No

Table 1. Minimum and maximum currency exchange rates in G-Econ for several countries.

G-Econ provides data on economic output in three different metrics: local currency [GCPLC], US$ according to market exchange rates [GCPMER], and US$ according to purchase power parities [GCPPPP]. Because currency exchange rates and purchase power parity data are determined at the national level, the ratios of GCPMER to GCPLC and of GCPPPP to GCPLC should be the same across all grid cells within a country. Contrary to this assumption, currency exchange rates in G-Econ 1.3 vary significantly across grid cells for some countries (see Table 1). This inconsistency strongly affects economic data for the two largest national economies, China and the USA, where exchange rates vary by several orders of magnitude.[21] The data inconsistencies for these two countries can be resolved by including data from the respective country files underlying the G-Econ database. For the USA, G-Econ+ uses GCPMER and GCPPPP data from the respective country file rather than from G-Econ. The country file for China contains a single value for GCPLC but two different values for GCPMER and GCPPPP. G-Econ uses the data from variant "B" in the country file, which are often grossly unrealistic[22] whereas G-Econ+ uses GCPMER and GCPPPP data from variant "A", which are based on reasonable exchange rates. The reasons for the inconsistent exchange rates in the other countries could not be resolved. The problem may be related to different treatment of output from the extraction of oil and mineral resources but it also affects grid cells and countries for which no separate information on economic output from oil production is available (see Table 1).

Data problems in G-Econ (global file):

- Some grid cells with non-zero GCP (gross product) have zero RIG (i.e., area).
- Population in areas with very low population density (according to GPW v3) has been set to zero in G-Econ, thus aggravating the problem of excess zeros in data aggregation and regressions.
- GCPMER and GCPPPP are partly or completely wrong for China, the USA, Angola, Kuwait, Yemen, Syria, Angola, Iran, Egypt, Saudi Arabia, and Libya, as indicated by varying MER and PPP exchange rates within a country (see below).
- GCPPPP is partly or completely wrong for Algeria, Botswana, Norway, and Venezuela, as indicated by varying PPP exchange rates within a country.
- GCPNOMER and GCPOILMER are unavailable for most grid cells. Furthermore, they are neither representative (e.g., the GCPOILMER data from the country file for Saudi Arabia is contained in some grid cells of G-Econ but not in others) nor consistent (e.g., GCPNOMER and GCPOILMER often do not add up to GCPMER, and GCPOILMER often exceeds GCPMER).

[21] W. Nordhaus acknowledges problems with the economic data for China and US but does not provide a clear explanation: *"We believe that the numbers for China and US in the data base have substantial errors […] I do not know what happened or why"* (personal communication).

[22] E.g., grid cell Lat=+35°, Long =+95°: GCPLC is ~12 Mio Yuan, GCPMER ("A") is ~2.6 Mio US$, GCPMER ("B") is ~155 Mio US$, GCPPPP ("A") is ~11.5 Mio US$, GCPPPP ("B") is ~670 Mio US$.

Data problems in the country files:

- GRID_AREA is often incorrect (e.g., Chile, Ukraine, Tanzania, Zimbabwe, Greenland, and Japan).
- POP is handled inconsistently. For most countries, country by cell population corresponds to the sum of all POP entries. For 16 European countries, however, country by cell population corresponds to the average of all (identical) POP entries.
- POP values for some subunits appear unrealistic (e.g., 26 persons for "West and South of Northern Ireland").
- GCPMER and GCPPPP appear reasonable when aggregated to the level of cell by country (but not necessarily at the level of cell by subunit). For most European countries, however, total GCPMER and GCPPPP at the level of cell by country is listed in a single arbitrary subunit. Furthermore, GCPMER and GCPPPP entries are zero for some countries.
- Market exchange rates and PPP exchange rates vary within Kuwait, Iran, Egypt, and Saudi Arabia.
- PPP factors vary within Venezuela.
- RIG is incorrect in Australia.

Inconsistencies between G-Econ and the country files:

- "Rate in grid" (RIG) from G-Econ and the country files are inconsistent for Australia and Angola. Furthermore, islands and inland lakes are often treated differently in G-Econ and the country files.
- Climate data from G-Econ and the country files differ substantially in some regions with steep climate gradients.
- Population data in G-Econ and the country files differ substantially in some coastal and border regions.
- Data on gross product in local currency from G-Econ and the country files differ significantly in many cases. For most African countries, the ratio of gross product in local currency between G-Econ and the country files is constant for all subunits in a country. Hence, the difference may be related to the choice of different base years for the currency conversion. For Turkey and Kuwait, however, this ratio differs widely across grid cells.
- GCPMER and GCPPPP from G-Econ and the country files (summed up to the level of cell by country) differ substantially for Turkey.

Acknowledgements

I am grateful to W.D. Nordhaus for making all G-Econ data publicly available and for his willingness to openly discuss my questions on G-Econ at the 3rd Atlantic Workshop on Energy and Environmental Economics (A Toxa, Spain, 4-5 July 2008) where I presented initial results of a statistical analysis based on the G-Econ+ dataset described here.