The influence of climate on global population distribution

Paper presented at the XXVI International Population Conference (IUSSP 2009), Marrakech, Morocco, 27 September – 2 October 2009

Hans-Martin Füssel
Potsdam Institute for Climate Impact Research
E-Mail: **fuessel@pik-potsdam.de**

Abstract

This paper assesses the influence of climatic factors on population distribution applying cross-sectional analysis to an updated version of the G-Econ database. Population density assumes a maximum for temperature in the 10°C to 14°C range and annual precipitation in the 1000 mm to 1400 mm range, with slight variations across central estimators. Temperature has a stronger influence on population density than precipitation. In particular, very cold temperature is a more severe constraint on population density than very warm temperature or insufficient precipitation. The influence of temperature and precipitation on population density is non-monotonous but the non-monotonicity decreases if the covariance between temperature and precipitation is taken into account. Population density exhibits large variability for a given climate, in particular in warm regions. Regression of population density on climatic factors is complicated by the presence of excess zeros, by the large variation of population density within temperature and precipitation bins, by the non-monotonous effect of temperature and precipitation on population density, and by substantial differences in the shape of the population distribution across central estimators. Therefore, it does not seem feasible to quantitatively estimate the impacts of anthropogenic climate change on future population distribution based on the results of this cross-sectional analysis.

1. Introduction

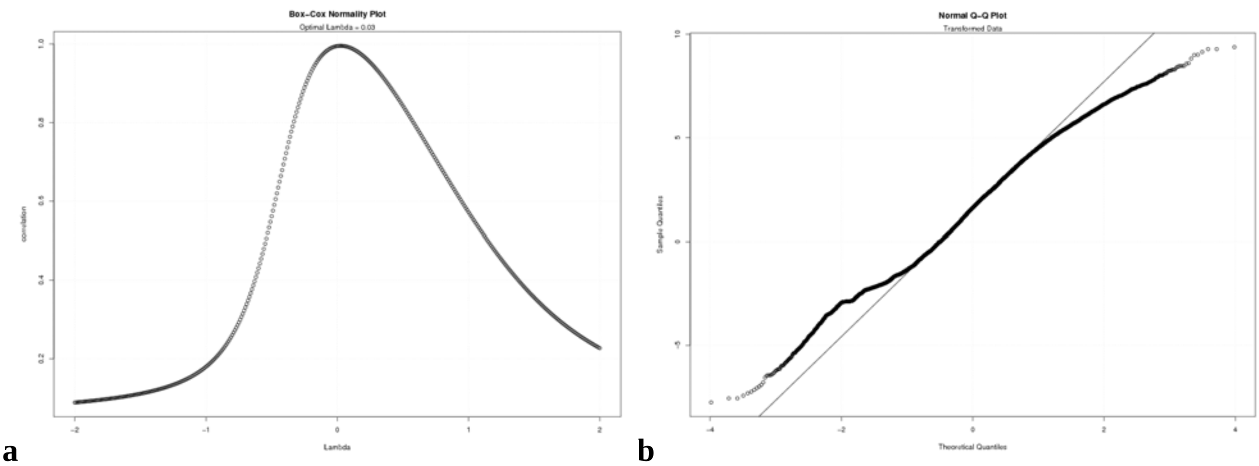
The population density and distribution in a region is determined by a complex interplay of geographic, climatic, environmental, historical, and cultural factors. Recognition of the severity of anthropogenic climate change has motivated research on the influence of climate on the current distribution of wealth. 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 [Mendelsohn and Schlesinger 1999, Nordhaus 2006, Nordhaus 2008, Dell et al. 2008, Dell et al. 2009]. The data collected in some of these studies are potentially useful for population studies as well.

G-Econ 1.3 presents climatic, geographic, demographic and economic variables for all 1°-by-1° terrestrial grid cells of the world [Nordhaus 2006]. The climate data in G-Econ 1.3 are based on direct observations and interpolations in data-poor regions [New et al. 2002]; the population data are from a variety of sources, including census data and on extrapolations from night-time lights [Balk and Yetman 2004]. This paper applies an updated version of the G-Econ database for a cross-sectional analysis of the relationship between climatic factors and population distribution. G-Econ+ corrects various flaws in the population and economic data of the original G-Econ 1.3 database (see the Appendix for details). G-Econ+ is available at two spatial resolutions: (i) all 17940 1°-by-1° land-based grid cells and (ii) for all 4103 national and subnational administrative units distinguished in G-Econ 1.3.

The following sections describe the methods applied, the individual influence of temperature and precipitation on population density, the relationship between temperature and precipitation, the climatic preferences of population, and the combined influence of temperature and precipitation on population density. This paper does not make projections of future population density under changed climate conditions.

2. Methods

The G-Econ+ database comprises climatic, demographic, and economic data for 17,940 terrestrial grid cells. 3,170 of these grid cells have zero population (i.e., the rate of zero entries is 17.6%). These "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. 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].



a b

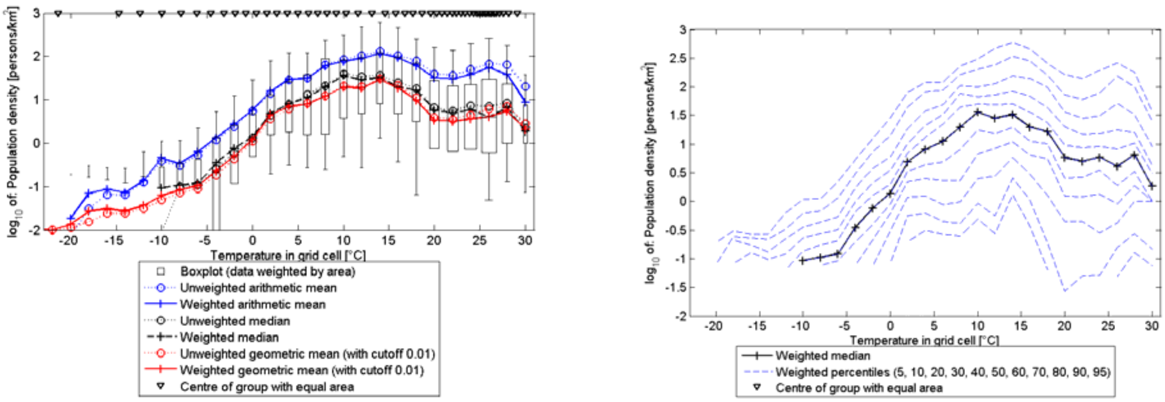
Figure 1: Box-Cox transformation for population density: (a) degree of normality over the transformation parameter lambda; (b) normal quantile-quantile plot of transformed variable.

Population data in G-Econ+ exhibits strong skewness and heteroskedasticity. Figure 1 presents the results of a Box-Cox transformation of population density, excluding cells with zero population are excluded. The optimal transformation parameter is λ=-0.03, which is very close to a logarithmic transformation. If excess zeros are neglected, a logarithmic transformation can make data on population density normal distribution-like. Any regression of population density on climatic and/or geographical variables, however, would have to account for excess zeros and correct the effects of the log transformation on central estimates of population density [Manning 1998].

Most analyses in the following sections depict the univariate or bivariate distribution of population density over temperature and/or precipitation. Grid cells are assigned to discrete temperature and precipitation bins, and population density is aggregated across all grid cells within a climate bin. The aggregation uses different central estimators: area-weighted and unweighted arithmetic mean, area-weighted and unweighted median, and area-weighted and unweighted geometric mean. Area-weighting considers that the grid cells in G-Econ have different land area. The total area of a 1°-by-1° grid cell decreases with absolute latitude, and coastal grid cells have a smaller land area than inland grid cells. The arithmetic mean represents average population density in the whole region covered by a particular climate bin; it is very sensitive to the presence of a few grid cells with very high population density (i.e., urban centres). The median represents population density in an average grid cell within a particular climate bin; it is insensitive to the presence of a few grid cells with very high population density as long as they represent less than half of the total area. The arithmetic mean is generally much larger than the median because population density is strongly skewed. The geometric mean corresponds to the arithmetic mean of log-transformed population density, which is approximately normally distributed if zero values are excluded. Due to the presence of excess zeros, however, the geometric mean of population density is highly sensitive to the choice of the lower cutoff point and it scales positively with the size of grid cells [Füssel 2009].

In addition to the central estimators, box plots and quantile plots are used to present further information on the distribution of population density within climate bins. The box plots in this paper denote the 25th, 50th and 75th percentile; the whiskers denote the 5th and 95th percentile.

3. Influence of temperature on population density



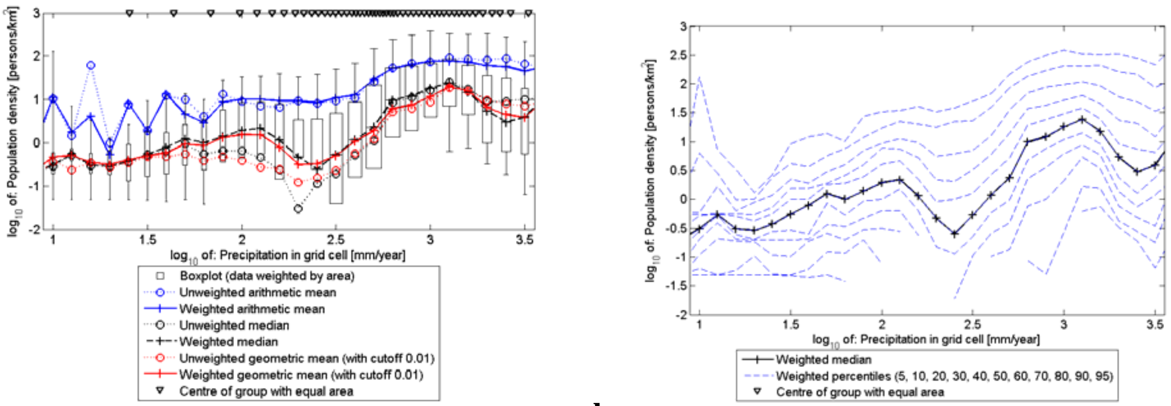
a **b**

Figure 2: Relationship between temperature and population density: (a) box plot; (b) quantile plot.

Figure 2 depicts the relationship between annual mean temperature (categorized in 2°C bins) and population density. Figure 2.a shows several central estimators and a box plot; the quantile plot in Figure 2.b shows the distribution of several area-weighted percentiles of population density. The main findings are as follows:

- The distribution of population density over temperature is non-monotonous and bimodal. Population density assumes a global maximum in the 10°C to 14°C temperature range, depending on the central estimator, and a local maximum in the 26°C to 28°C range.
- The arithmetic mean and the median of population density have a similar distribution across temperature but the value of the former is much larger (factor 3 to 10) than that of the latter. With a cutoff point of 0.01 persons per km^2, the geometric mean of population density is very similar to the median where the latter is defined.
- Population density is highly variable within temperature bins. The interquartile range spans more than one order of magnitude; it is smallest around 12°C and around 0°C, and it becomes infinite (or undefined) below -4°C.
- The distribution of population density across temperature is rather insensitive to area-weighting of grid cells.
- The area of unpopulated grid cells comprises less than 10% of the total area in all temperature bins between 0°C and 30°C but more than 50% of the total area in all temperature bins below -10°C. Hence, very low temperature is a strong constraint for (even sparse) human population of a region.
- Temperature has a qualitatively similar effect on all percentiles of population density. The effect on log-transformed population density is largest for the lower percentiles.

4. Influence of precipitation on population density



a **b**

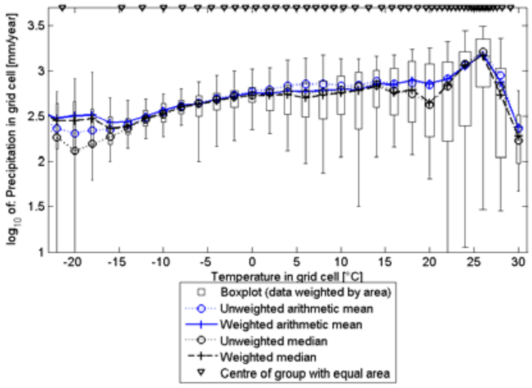
Figure 3: Relationship between precipitation and population density: (a) box plot; (b) quantile plot.

Figure 3 depicts the relationship between annual precipitation (log-scaled; i.e., the value 3 on the x axis denotes 1000 mm per year) and population density. The main findings are as follows:

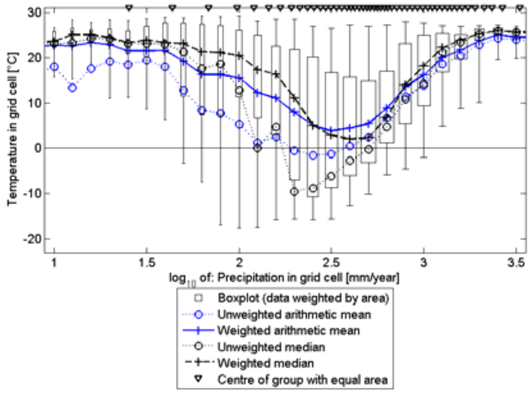
- The distribution of the median of population density over precipitation is non-monotonous whereas the distribution of the arithmetic mean is monotonous across most of the precipitation range. For annual precipitation above 250 mm, both central estimators of population density increase up to a global maximum in the 1000 to 1400 mm range; the decrease at higher precipitation levels is more pronounced for the median than for the arithmetic mean.
- The arithmetic mean of population density is much larger (factor 2 to 30) than the median. Outside the 250 mm to 1400 mm precipitation range, the distribution of population density differs qualitatively between these two central estimators. With a cutoff point of 0.01 persons per km^2, the geometric mean of population density is similar to the median.
- The distribution of population density across precipitation is more complex than across temperature. Above 100 mm precipitation, the median varies much more across precipitation levels than the arithmetic mean due to the presence of large population centres even in precipitation bins where average grid cells have low population density. Below 100 mm precipitation, the arithmetic mean varies much more across precipitation levels than the median, suggesting that large population centres are present in some but not all of the very dry precipitation bins.
- Population density is highly variable within precipitation bins. Neglecting very dry regions with annual precipitation below 50 mm, the interquartile range generally spans more than one order of magnitude and assumes a minimum around 1200 mm per year.
- The distribution of population density across precipitation can be very sensitive to area-weighting of grid cells. Between 100 mm and 500 mm annual precipitation, the unweighted median is much lower (factor 2 to 10) than the weighted median. This result can be explained by the strong presence of very cold, sparsely populated high-latitude grid cells with small land area that represent half of the grid cells but less than half of the area in this precipitation range (see Figure 4 below).
- The area of unpopulated grid cells comprises less than 10% of the total area in all precipitation bins below 100 mm and above 500 mm but more than 20% of the area in precipitation bins between150 mm and 350 mm. These results can only be explained if the covariance between precipitation and temperature is taken into account (see Figure 4 below).

- Precipitation has qualitatively different effects on different percentiles of population density. The upper percentiles assume a minimum in very dry regions with annual precipitation around 20 mm whereas the lower percentiles assume a minimum in moderately dry regions with precipitation between 100 mm and 500 mm.

5. Relationship between temperature and precipitation



a



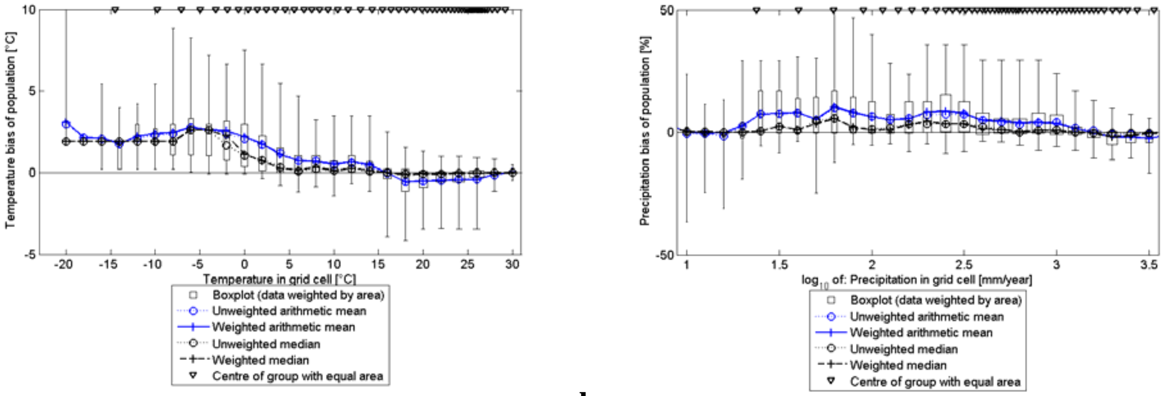
b

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

Mean population density exhibits a local maximum around 26°C (Figure 2) and the median of population density exhibits a global minimum around 250 mm precipitation (Figure 3). To understand the reasons for these extrema, Figure 4 explores the covariance between temperature and precipitation. The main findings are as follows:

- The relationship between annual temperature and precipitation is non-monotonous.
- Mean precipitation increases slightly with temperature up to around 26°C, and decreases substantially for even warmer temperature. This result suggests that the local maximum in population density around 26°C is primarily caused by very high precipitation rather than by particularly favourable temperatures.
- Very cold regions (below 0°C) make up more than half of the grid cells and about half of the area in precipitation bins between 200 mm and 500 mm per year. This result suggests that the median of population density assumes a minimum in this precipitation range mainly because of very low temperature rather than particularly unfavourable precipitation levels.
- Very low and very high precipitation occurs almost exclusively in warm regions with a temperature above 20°C; colder regions have more moderate precipitation levels (except the very coldest regions below -18°C, which can also be very dry).

6. Climatic preference of population



a b

Figure 5: Climatic preference of population within administrative units (see text for details): (a) Temperature bias; (b) precipitation bias.

The availability of G-Econ+ at two different spatial resolutions allows for an independent analysis of the influence of climatic factors on regional population distribution. Figure 5 depicts the temperature and precipitation bias of population within administrative units, represented by the difference between population-weighted and area-weighted temperature and precipitation, respectively. Note that the magnitude of the climate bias across temperature and precipitation regimes should be interpreted with great 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 temperature and precipitation. The main findings of Figure 5 are as follows:

- The temperature bias of population is strongly positive below about 2°C, slightly positive up to 16°C, and slightly negative above 16°C. In other words, population density is generally higher in the warm parts than in the cold parts of a cold administrative unit.
- The precipitation bias of population is slightly positive up to about 1000 mm and neutral above that level. In other words, population density is generally higher in the wetter parts than in the drier parts of an administrative unit except for very humid regions.
- The local extrema of population density identified in Figure 2 and Figure 3 are not detected here, which provides further evidence that they are mainly due to the significant covariance between temperature and precipitation.

7. Bivariate distribution of population density

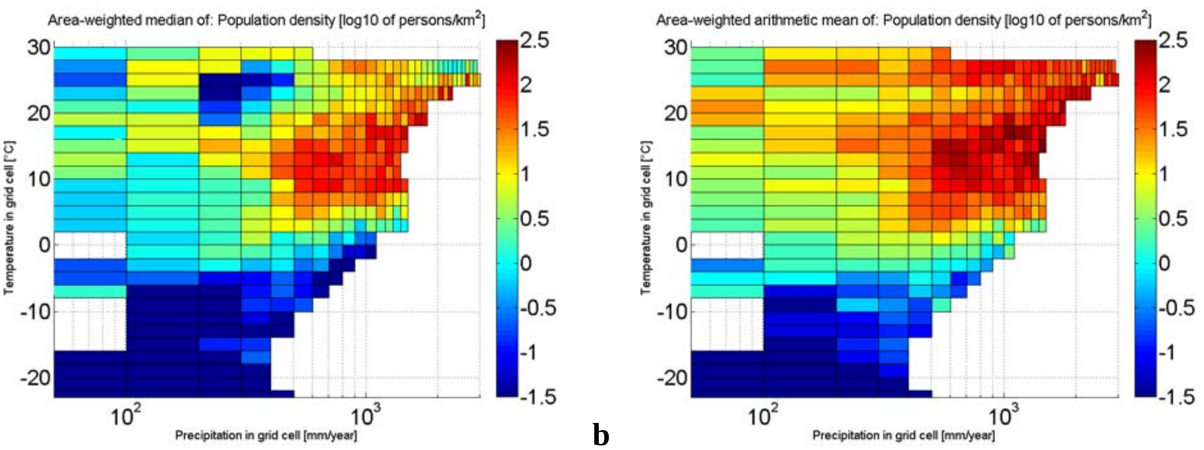


Figure 6: Combined influence of temperature and precipitation on (a) median and (b) arithmetic mean of population density (colour-coded).

The strong non-monotonous relationship between temperature and precipitation shown in Figure 4 confounds the univariate distribution of population density over these climatic variables. In order to separate their effects, Figure 6 depicts the bivariate distribution of population density over temperature and precipitation. The main findings are as follows:

- The median of population density is largest in areas with annual precipitation above 500 mm and with temperature between 6°C and 20°C; this range extends to 28°C for the arithmetic mean. The difference between the two central estimators in the 24°C to 28°C range was already observed in Figure 2. Hence, most regions in this temperature regime are sparsely populated but there are also many large population centres that can be found in all but the driest regions.
- Population density generally increases with precipitation across the full temperature range.
- For most precipitation levels, the relationship between temperature and population density is inverse U-shaped for the median and almost monotonically increasing for the arithmetic mean.
- Unfavourable (i.e., very low) temperature is a much more severe constraint to the population of regions than unfavourable (i.e., very low) precipitation.
- Optimal precipitation levels are positively correlated with temperature. For instance, annual precipitation of 600 mm is sufficient for high population density in most regions of the 6°C to 16°C temperature range but not above that temperature range.

8. Summary and conclusions

This paper has assessed the influence of climatic factors on population distribution applying cross-sectional analysis to an updated version of the G-Econ database. The main results of this analysis are as follows:

- Population density assumes a maximum for temperature in the 10°C to 14°C range and annual precipitation in the 1000 mm to 1400 mm range, with slight variations across central estimators.
- Temperature has a stronger influence on population density than precipitation. In particular, very cold temperature is a more severe constraint on population density than very warm temperature or insufficient precipitation.
- The influence of temperature and precipitation on population density is non-monotonous; this non-monotonicity is much stronger for the median than for the arithmetic mean. The non-monotonicity is partly caused by the covariance between temperature and precipitation. For a given precipitation or temperature level, the arithmetic mean of population density generally increases with increasing temperature and precipitation, respectively. This statement does not hold for the median.
- Population density varies strongly for a given climate. The interquartile range generally comprises more than an order of magnitude, and it is particularly large in warm regions.
- The arithmetic mean of population density for a given climate is much larger (factor 2 to 30) than the median (and the geometric mean) because of strong skewness.

This analysis has shed much light on the influence of climatic factors on current population distribution. Regression of population density on climatic factors is complicated, however, by the presence of excess zeros, by the large variation of population density within temperature and precipitation bins, by the non-monotonous effect of temperature and precipitation on population density, and by substantial differences in the shape of the population distribution across central estimators. Therefore, it does not seem feasible to extrapolate the results of this analysis to quantitatively estimate the impacts of anthropogenic climate change on future population distribution.

9. References

[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.

[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.

[Füssel 2009] Hans-Martin Füssel. New results on the influence of climate on the distribution of population and economic activity. MPRA Paper No. 13788, 2009. URL http://mpra.ub.uni-muenchen.de/13788/.

[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.

[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.

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

[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.

[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*, 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.

[Welsh et al. 1996] A. H. Welsh, R. B. Cunningham, C. F. Donnelly, and D. B. 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

A1. Introduction

G-Econ 1.3 provides data on climatic and geographic factors, population, and economic output[1] 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[2] 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[3] 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"[4] but it does not contain information on the 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.

A2. 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

[1] Analogous to Nordhaus, the terms "economic output" and "gross product" are used interchangeably here.
[2] 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).
[3] A Toxa, Spain, 4-5 July 2008, http://webs.uvigo.es/rede/toxa/pages/3rd-atlantic-workshop/program-papers.php
[4] 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.

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+

A3. 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[5] according to the "Rate in grid" (RIG) data from the from country files.

[5] 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.

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 0 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.

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[6] 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[7], the grid cell was excluded from

[6] 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+.

[7] 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.

the calculation of average output per capita. In addition, GDP per capita for Alaska was set to the average value of the USA.[8] An analogous procedure was applied for gross product based on purchase power parities (PPP) rather than market exchange rates (MER).

A4. 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
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

[8] 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.

(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.[9] 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[10] 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).

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.

---

[9] 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).

[10] 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$, GCPPPP ("B") is ~670 Mio US$.

- 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.