ANNEX C

Indexes and estimation techniques

Theil entropy index

Definition: Regional disparities are also measured by a Theil entropy index, which is defined as:

$$Theil = \sum_{i=1}^{N} \frac{y_i}{\overline{y}} ln\left(\frac{y_i}{\overline{y}}\right)$$

where N is the number of regions in the OECD, y_i is the variable of interest in the i-th region (i.e. household income, life expectancy, homicide rate, etc.) and $\overline{y}$ is the mean of the variable of interest across all regions.

The Theil index can be easily decomposed in two components: i) the disparities within subgroups of regions – where for example a subgroup is identified by a set of regions belonging to a country; ii) the disparities between subgroups of regions (i.e. between countries). The sum of these two components is equal to the Theil index.

In order to decompose the Theil index, let us start by assuming m groups of regions (countries). The decomposition will assume the following form:

$$Theil = \sum_{j=1}^{M} \sum_{i=1}^{N} s_j \frac{y_{ij}}{\overline{y}_j} ln\left(\frac{y_{ij}}{\overline{y}_j}\right) + \sum_{j=1}^{M} s_j ln\left(\frac{\overline{y}_j}{\overline{y}}\right)$$

where the first term of the formula is the *within* part of the decomposition equal to the weighted average of the Theil inequality indexes of each country. Weights, s_j, are computed as the ratio between the country average of the variable of interest and the OECD average of the same variable. The second term is the between component of the Theil index and represents the share of regional disparities that depends on the disparities across countries.

Interpretation: The Theil index ranges between zero and ∞, with zero representing an equal distribution and higher values representing a higher level of inequality.

The index assigns equal weight to each region regardless of its size; therefore, differences in the values of the index among countries may be partially due to differences in the average size of regions in each country.

Methodology to estimate the potential for remote working

The assessment of regions' capacity to adapt to remote working is based on the diversity of tasks performed in different types of occupations and is structured in two steps.

The first step requires classifying each occupation based on the tasks required and according to the degree to which those tasks can be performed remotely. Such a classification is based on a recent study by Dingel and Neiman (2020, "How many jobs can be done at home?", Becker Friedman Institute White Paper March, *https://bfi.uchicago.edu/working-paper/how-many-jobs-can-be-done-at-home/*), which is built from the O*NET surveys conducted in the United States. The second step relies on data from labour force surveys and consists of assessing the geographical distribution of different types of occupations and subsequently matching those occupations with the classification performed

in the first step. Combining the two data sets allows assessing the number of workers that can perform their task from home as a share of the total employment in the region.

This assessment does not consider the specific regulations or arrangements that each country applies to remote working and which affect the actual share of people working remotely. For example, limitations in the days of remote working for cross-border workers are not reflected in the estimates presented here.

Methodology to estimate GDP at the metropolitan level

The proposed methodology uses GDP per capita values in TL3 regions, TL2 regions (for Australia, Chile, Colombia and Mexico) and census metropolitan areas (CMA) in Canada as data inputs and the distribution of the population based on the Global Human Settlement (GHS) population grids.

The suggested methodology is composed of three main steps:

1. Intersect the functional urban area's (FUA) boundaries with the TL3 boundaries.
2. Calculate the share of population living in the intersection of the TL3 boundary and the FUA.
3. Derive the gross domestic product (GDP) value based on the share of population living in the area calculated in the previous step.

It has to be noted that the estimates of GDP in the metropolitan areas do not adhere to international standards; the comparability among countries relies on the use of the same methodology applied to areas defined with the same criteria.

For the United States, county-level data was aggregated to FUAs.

Methodology to estimate cooling degree days at the FUA level

The data used to compute cooling degree day (CDD) indicators at the FUA level comes from the historical global gridded degree days database of CDD and heating degree days (HDD)The database includes three types of indicators corresponding to CDD, HDD, and CDD computed using wet-bulb temperature (CDDwb). Each indicator is available at 6 different threshold temperatures: 18, 18.3, 22, 23, 24 and 25°C for CDD and CDDwb and 10, 15, 15.5, 16, 17 and 18°C for HDD. The database provides these three indicators both by year and by month over the period 1970-2018.

The dataset used to compute indicators at the FUA level is the CDD raster corresponding to a threshold temperature of 22°C. The 49 bands of the raster correspond to the annual CDD values from 1970 to 2018 included.

Indicators were computed using the geopandas, rasterstats python libraries and by intersecting the raster file with the shapefile corresponding to the FUAs' boundaries. For each FUA, the average cell value is calculated. All cells having an intersection with the FUA are included in the mean value calculation. The cells with missing values are ignored.

Methodology to estimate electricity indicators at the regional level

The Global Power Plant Database (GPPD) provides information on power plants located in 164 countries all over the world, including the 37 OECD countries. For each power plant, the GPPD provides the geographic co-ordinates and a number of attributes, as follows:

- The energy source: oil, gas, coal, petroleum coke, cogeneration, hydro, wind, waste, biomass, wave and tidal, geothermal, solar, nuclear and others.
- The generation capacity, which is the maximum power (in megawatts, MW) that the plant can deliver. The capacity is a facility-specific characteristic and does not change over time, unless extension or upgrade of the power station, or a shutdown of a part of it.

- The annual electricity generation, which provides the amount of electricity generated over a year (in GWh). This indicator is reported for 24% of the power plants over the period 2013-17. When no electricity generation was reported, the annual electricity generation was estimated. The annual generation corresponds to the gross generation, i.e. the electricity consumption of the power plant for its operation is not deducted.
- The country where the power plant is registered.

The International Energy Agency (IEA) database (see Annex B) includes national-level electricity generation data by energy source for most OECD countries (excluding Colombia). The IEA dataset used to estimate electricity generation indicators at the local level corresponds to the gross electricity production by energy source in 2017. A breakdown of 53 different sources is available.

Electricity generation estimates

In order to remain consistent across countries and energy sources, electricity generation was estimated at the power plant level based on the relative capacity of each power plant (from the GPPD) and on the total national electricity generation form each energy source (from the IEA). The methodology follows the four steps below:

1. Map energy sources from the IEA to the GPPD classification.
   The IEA electricity production data provides a higher level of detail in terms of breakdown by energy source compared to the GPPD data. For this reason, each energy source type recorded in the IEA database was matched to a source category in the GPPD.
2. Determine the share of national capacity for each power plant.
   For each power plant *p*, located in the country *c* and generating electricity from the energy source *f*, the share of the capacity of the power plant in the national capacity for the source *f* is calculated as:

$$share_{p,c,f} = \frac{capacity_{p,c,f}}{\sum_i capacity_{i,c,f}}$$

where $i \in$ power plants located in the country *c*, and generating electricity from the source *f*.

1. Allocate a part of the national generation to each power plant.
   For each power plant *p*, generating electricity from source *f*, in the country *c*, the estimated generation is calculated as:

$$generation_{p,c,f} = share_{p,c,f} * national\ generation_{c,f}$$

1. Exceptions.
   Since no data on electricity generation by source is available for Colombia in the IEA database, only the GPPD estimated generation data was used. In contrast, GPPD data was not necessary to estimate electricity production within Estonia, Latvia and Luxembourg, as those countries do not have a geographical disaggregation according to the OECD definition of large regions (TL2) (see the OECD Territorial Grid in Annex A). For these countries, direct use of IEA was sufficient for comparisons with other TL2 regions.

Aggregation at local scales

In order to compute indicators at different geographical scales, a point shapefile was created from the GPPD using the latitude and longitude provided for each facility – each point representing a power plant. The point shapefile was overlapped with two other shapefiles corresponding to the boundaries of the subnational geographies available in OECD countries (TL2 and TL3 regions). Thus, each power plant can be associated to a TL2 region and a TL3 region. Offshore power plants were assigned to the closest region (of the registered host country) based on the distance to the coast.

Year of reference

All indicators presented in this document refer to the year 2017, which corresponds to the latest year for which capacity data is available in the GPPD.

Breakdown by energy source categories

The GPPD includes 13 different energy sources. These energy sources were aggregated into three categories (fossil fuels, renewables and nuclear). The energy sources within each category are comparable in terms of technology, risks and impacts on the environment. A sub-category for coal was made, as coal is the most carbon-intensive fuel to produce electricity.

Electricity generation indicators

For each region *r*, generation data was aggregated into each category *i* as:

$$generation_{r,\ i} = \sum_{k\ \in\ i} power\ plant\ generation_{r,k}$$

where *k* ∈ {coal, gas, oil, petroleum coke, cogeneration, nuclear, hydro, wind, waste, biomass, wave, geothermal, solar}, *i* ∈ {fossil fuels, coal, nuclear, renewables}, and $power\ plant\ generation_{r,\ k}$ is the electricity generation of a power plant located in the region *r*, generating electricity from the source type *k*.

Energy mix indicators

For each region *r*, the share of each energy source category *i* (fossil fuels, coal, nuclear, renewables) is calculated as:

$$share_{r,i} = \frac{generation_{r,\ i}}{\sum_j generation_{r,\ j}} * 100$$

where j ∈ {fossil fuels, renewables, hydro, wind, nuclear}.

Greenhouse gas (GHG) emissions from electricity generation indicators

GHG emissions indicators are derived from both the electricity generation by energy source and the emission intensity of each energy source. Electricity generation was estimated at the power plant level for each energy source included in the GPPD as described above. Emission intensity by energy source comes from the IPPC estimates on GHG emissions of supply technologies.

For each region *r*, the GHG emissions (in tons of CO_2 equivalent) are calculated as:

$$emissions_r = \sum_{k\ \in\ f} generation_{r,\ k} * emission\ intensity_k$$

where the emission intensity corresponds to the median value of the lifecycle emissions (in gCO_2eq/kWh), f ∈ {coal, gas, oil, petroleum coke, cogeneration, nuclear, hydro, wind, waste, biomass, wave, geothermal, solar}.

Emission intensity indicator

For each region r, the emission intensity (in tons of CO_2 equivalent per GWh) is calculated as:

$$emission\ intensity_r = \frac{emissions_r}{\sum_i generation_{r,i}}$$

where *i* ∈ {fossil fuels, renewables, nuclear}.

Methodology to estimate protected areas at the regional level

The World Database on Protected Areas (Annex B) is a worldwide record of marine and terrestrial protected areas. Launched by the International Union for Conservation of Nature (IUCN) and the United Nations Environment Programme (UNEP), the geospatial database has been compiled and is

updated monthly by the UN Environment Programme World Conservation Monitoring Centre (UNEP-WCMC).

The database is made up of about 242 000 records of protected areas, split into 2 shapefiles. Each protected area is recorded either as a polygon, delimiting the boundaries of the area or as a point with a reported area providing information on the extent of the protected area. One shapefile contains all the protected areas recorded as polygons and the other one is for protected areas recorded as points.

Non-geospatial information is also available for each record, giving more details on the protected areas. Among the 28 fields accessible through the attributes table, 5 are useful for the analysis described in this document:

- IUCN management categories (IUCN_CAT): The different categories of protected areas made by the IUCN correspond to the management objectives within the areas. Seven different categories can be distinguished, going from the most restrictive natural zone management to a zone with sustainable use of natural resources (Ia: Strict Nature Reserve; Ib: Wilderness Area; II: National Park; III: Natural Monument or Feature; IV: Habitat/Species Management Area; V: Protected Landscape/Seascape; VI: Protected area with sustainable use of natural resources). This variable can also take the following values: not applicable, not assigned or not reported.
- Status (STATUS): Refers to the administrative status of the protected areas: “Designated”, “Inscribed”, “Adopted”, “Proposed” or “Established”.
- Status year (STATUS_YR): Year corresponding to the entry into force of the current status of the protected area.
- Designation (DESIG): Corresponds to the subnational, national or international framework or agreement the protected area is part of.
- Reported area (REP_AREA): Protected area extent (useful for protected areas recorded as points).

Following the methodology developed for country-level indicators (see Mackie, A., et al. (2017), «Indicators on Terrestrial and Marine Protected Areas : Methodology and Results for OECD and G20 countries», *OECD Environment Working Papers*, n° 126, Éditions OCDE, Paris, *https://doi.org/10.1787/e0796071-en*), protected areas with “not reported” or “proposed” status, and UNESCO Man and Biosphere Reserves are excluded for the analysis as well as protected areas recorded as points without a reported area.

The shapefile containing protected areas recorded as polygons was dissolved to avoid overlaps between protected areas and converted afterwards into a 300 meter-resolution raster file. The raster does not take into account IUCN management categories.

Two indicators (share of regional protected area and share of regional coastal protected area) are computed from this raster file, following the steps below:

1. Share of regional protected area
    - The regional area (RA) is calculated from the regions’ shapefile.
    - The regional protected area extent (PA) is calculated from the protected areas raster, the protected areas recorded as points shapefile and the regional boundaries’ shapefile. The first part of the regional protected area extent (PA1) is calculated as the sum of the reported areas of all the points located within the region. The second part (PA2) is calculated as the protected zones extent within the regional boundaries measured from the raster. The regional PA is thus calculated as PA1 + PA2.
    - The share of protected area within the region (%) is calculated as 100*PA/RA.

2. Share of regional coastal protected area
   - A 50 km-buffer is created around the coastlines.
   - The regional coastal area (CA) is calculated for each region as the area of the intersection between the 50 km-buffer and the regions' shapefile.
   - The coastal protected area extent (CPA) is calculated from the protected areas raster, the protected areas recorded as points shapefile, the 50 km-buffer and the regional boundaries' shapefile. The first part of the coastal protected area extent (CPA1) is calculated as the sum of the reported areas of all the points located within the intersection between the buffer and the region. The second part (CPA2) is calculated as the protected zones extent within the intersection between the buffer and the region measured from the raster. The CPA is thus calculated as CPA1 + CPA2.
   - The share of coastal protected area within the region (%) is calculated as 100*CPA/CA.