III. HOW TO LINK SPATIAL DATA TO POPULATION DATA

Linking population data to the geography of risk seems to be an easy task. However, it is made difficult because censuses publish or distribute information by administrative areas that may not coincide with environmental areas (see Balk and Yetman, 2004). In their article, "The Rising Tide: Assessing the Risks of Climate Change and Human Settlements in Low Elevation Coastal Zones", McGranahan, Balk and Anderson (2007) assessed the distribution of human settlements in Low Elevation Coastal Zones (LECZs) around the world. In order to calculate the population at risk and their international distribution in LECZs, the authors integrated spatially constructed global databases of population distribution, urban extent and elevation data, overlaying gridded geo graphic data, thus deriving totals of national populations in LECZs. Although the authors are able to calculate exposure of coastal areas to sea level rise, they also agreed that this is just a first appraisal and further disaggregation is needed.

As described before, survey data are typically rendered in fairly coarse spatial terms (subnational regions like provinces). This implies a fairly limited use when combined with spatial data on climate-risks such as low-lying coastal zones. In contrast, census data have much greater intrinsic spatial flexibility. Estimation of populations at risk has to date relied heavily of using population counts for small areal units (sometimes transformed as we describe below). These small areal units are the back bone of the census. Any census variable that is reported at the level of very fine areal units can be combined with geographic information system (GIS) tools into geographically identified regions. In some countries, only population counts are made available at the finest level, whereas additional variables may be available for coarser units. With very fine units, one will have many geographic tools and methods available to use to create estimates of population characteristics by various geographic zones.

Census data are typically not used as micro-data, to maintain confidentiality, but rather as attributes of small administrative units. This type of aggregation – while very powerful in some respects – can also lead to misleading conclusions. For example, if one were to find that population of the coastal zone was more urban than the population living outside the coastal zone, and that the population of the coastal zone is also wealthier than the population living outside the coastal zone, one

might infer that urban dwellers are wealthier than others. But this type of inference arises through an ecological fallacy (REF) and is quite possibly false. To properly assess this inference, one would need to return to the micro data, identify urban households at different wealth levels and rural households at different wealth levels, then re-aggregate to the small administrative units. (Similarly, when survey data are used in this way as summaries at the subnational regional level, they too fall prey to potential ecological fallacy.)

This raises an important point about preparing census data for environment and climate analysis. Because the end result will inevitably be geographic units, aggregation from micro data will always occur. This aggregation will always leave analysis open to the ecological fallacy, unless the analyst selects and constructs the right crosstabs in the micro data. This means that an analysis of the dwelling type and service access of women headed households must begin with aggregation of micro data on dwelling type and service access by household headship. Once aggregation is done, it may be time consuming to go back to micro data, or even impossible depending on the nature of access, meaning that it is essential to carefully think through the crosstabs needed for the analysis ahead of time.

Because geographic zones apart from administrative units have not been commonly used in the past, census analysts have not prepared summaries of census data for those zones from the micro-data themselves. Existing technology makes the construction of different geographies aggregated from micro data very possible within the confines of the national statistical offices or their protective enclaves (electronic or otherwise), though at present time this is not common practice.

What does it mean to make demographic data relevant to climate change?

Climate change is a spatial phenomenon. To make population data relevant, they must also be rendered spatially. This means that small area spatial units data and key indicators on population distribution and composition are both necessary. Spatial data formats vary and the tools for working with them also vary accordingly. Administrative data are typically vector-format polygons. (See Box 1.) Once population data are rendered in small spatial units (enumeration areas, blocks, etc.), it is important that they be integrated with spatially-specific climate change

data. Climate data are almost always raster-format or grids. Some form of correspondence between any two spatial units that are not identical is required. When linking population data with climate data, that integration takes places in a spatial framework, and depending on what is being integrated, may require that population data are transformed from irregularly-shaped census units (usually, in vector format) to a uniform grid (or raster format). Transformation to a grid helps reduce data loss and facilitates consistency in the generation of estimates. Box 2 briefly describes how these transformations take places and one common assumption used and the rules of proportional allocation, to accomplish it.

Alternatively, it may involve summarizing climate data in raster form according to small area polygons of the administrative data. A key decision point in this type of analysis involves when to use a grid as the basis of analysis, and when to use polygons like enumerator areas or administrative units as the basis of analysis. Some guidelines for this decision are as follows:

- *Enumerator areas or administrative boundaries often are constructed to have social meaning, be they neighbourhoods, blocks, communities, municipalities, provinces or even national boundaries. Depending on the reasons for the analysis, it may be important to retain this social meaning in the results. In these circumstances, it is best to retain the census unit as the base.*
- *Sometimes several different types of geography are important to the analysis; for instance, water catchment areas, urban boundaries, flood plains and low elevation coastal zones. In these instances, it may be more important to be able to move between these geographies quickly and easily, and transforming the population data to a grid is likely to be the best choice.*
- *Comparisons between censuses with different units – either across years when the units have been changed, or between countries – require a consistency of unit that cannot be delivered by polygons that do not match in size or time. Under these circumstances, transformation to a grid provides the most consistent base for analysis.*

Using a polygon base ultimately works in much the same way as transforming to a grid when either developing new geographies or

summarizing statistics from raster to polygon. Box 1 describes the proportional allocation method. Similarly, raster areas can be outlined and transformed to polygons, and these polygons can be matched with census polygons to determine the extent of overlap, which in turn determines allocation.

Alternatively, analysts can maintain the census unit, and then identify the average value of a set of pixels within a particular polygon quite easily in standard mapping software (often referred to as zonal statistics). This is common for remotely observed variables like temperature. In the end, if the result is to be a level of exposure for a neighbourhood, zonal statistics to summarize raster-form climate data are the best bet. If the result is to be the population size and composition of a geography of risk, either gridding or maintaining the census unit may work.

Coastal population distribution provides a strong example of the benefits of gridding. To date, among the many climate-related or environmental risks to population, only coastal population distribution has been systematically estimated in an integrated fashion. Until recently coastal proximity was not a consideration in demographic analysis and for the United States for example, a country with much flexibility in how it could repackage its demographic estimates, the initial estimates of coastal population has been greeted with some skepticism (Crowell et al., 2007).

Fortunately, increasing data availability and development of new estimation methods over the past decade are making estimation possible even in low-income countries. One of the first studies to systematically identify global population distribution with respect to coastal proximity was that of Small and Cohen (2004). They defined coastal proximity as residence 'within 100 km' of a coastline, this distance being the best that could be done at the time, given the coarse spatial resolution of the population data then available. Small and Cohen found that one-third of the global population lives within 100 km of a coast. Small and Nichols (2003), in addition to describing population distribution, found that the coastal population lives at densities at least three times that of population farther from the coast. A more recent study by McGranahan and colleagues (2007) employed more refined measures of coastal proximity, and drew upon data sources that distinguished urban from rural population and land areas. In that study, using elevation maps from

satellite data, coastal proximity was defined as the land area contiguous to the sea-coast up to a 10-meter level of elevation (i.e. the Low Elevation Coastal Zone, LECZ). This definition results in a coastal zone that varies in width from the coast-line. For example, a sea cliff more than 10 meters above sea level would not be included in the zone, whereas very low lying deltas might have land area in the zone out as far as 50 km or more from the coast line. The advances in the 2007 study were largely made possible by investments in finer resolution population data (used by the GRUMP project, CIESIN et al., 2004), and improvements in satellite measures of elevation that allow for refinements in estimates of coastal elevation.

But it is not only for coastal populations that demographers lack spatially detailed data – for poor countries this has been a limitation for all types of locations. Outside the high-income countries, which hold regular censuses and have statistical systems capable of collecting, mapping, and analyzing spatially-specific population data, very little is known of the demographic features of any population that does not conform to regular and usually, coarse reporting units. Nor it is a limitation that is easily overcome.

It can be quite difficult to convert population data organized by administrative units into estimates of population distribution across space. Census data are typically reported for administrative units such as provinces, states or in some cases municipalities. Usually these data are summarized in a database (or set of tables) that is organized by administrative names. Very often the spatial boundaries associated with these administrative units, even at this level of disaggregation, are not made publicly (or, at least not freely) available. Even within units of national statistical offices, data may not be available to all units within the agency. In many national statistical offices boundary data are the domain of geographers and population data are the domain of demographers and efforts to combine data are sometimes limited to the most basic, coarse-level reporting units. Even in countries where there are high-degrees of cooperation within the national statistical office, sometimes regional, state or province-level and city agencies and planners cannot access necessary data. Cooperation between national statistical offices and state and local agencies that either want to use census data or may create their own related data is highly recommended, particularly in the context of climate change. Adaptation will be local,

and many of the changes to the climate will be observed locally. Therefore, greater cooperation by like-minded agencies operating at different levels of government and administration is likely to beneficial to all agencies involved.

Even when spatial units that match census reporting data are available, the spatial and administrative data are seldom linked, leaving the data user to grapple with the challenges of manipulating and reconciling conventional tabular data with spatial data. Some specialized knowledge and training necessary to work with these different data effectively. This is an important problem, especially at the local level where expertise in the many areas required by interdisciplinary analysis would be hard to come by. Therefore, NSOs should make every effort to maintain linkages between disparate data types. For example, data tables of demographic characteristics that are organized by geographic regions should retain the name and complete code of that region. Similarly, geographic data should retain not only codes and names of the smallest possible unit, but the hierarchical information that allows smaller units to be matched to other administrative or political geographic units.

A methodological issue that is of particular concern for spatially defined areas such as coastal areas or flood zones is the spatial resolution of units. Note that these types of zones are not unique. Many ecologically defined zones are irregular and cut across many possible administrative units. The finer the unit of interest – for example, the finest grained units that might border a coastline or river – the more difficult are the data to acquire. This creates an inherent problem when the objective is to estimate population characteristics in a narrow geographic area such as a strip of coastal land. Even when the coastal band is sizable, its area will usually not generally conform to the formal boundaries of administrative units.

Using Vietnam[1] as an example, here it is shown why the resolution of population data matters for estimating populations facing coastal hazards – i.e., living in a low elevation coastal zone (LECZ) (McGranahan et al., 2007). Figure 1 shows a close-up of several first-order administrative units – provinces – in Vietnam. The finely detailed

---

[1] The examples for Vietnam were created with Veronique Marx of the UNFPA Vietnam field office.

boundaries shown – they are fourth-order administrative units – are termed communes, shown in Figure 2. Vietnam is unusual for a developing country in that the resolution of its spatial data is high. These data are fine-grained enough that the native data format (i.e., vector) may be overlaid with data on the LECZ to estimate the population living at risk of coastal hazards.

Overlaying data in this way brings a number of analytic problems to the fore. For any commune that intersects the LECZ (rather than being fully covered by it), an assumption must be made about how to estimate the population in that unit. For some purposes one might want to include the entire population in any administrative unit that intersects the LECZ. For example, municipal governments have limited resources. If flooding were to occur in a limited area the economic burden to prepare or respond to the flooding would in some part be shared by the entire population of a given municipality, so the population of the unit as a whole may be the best estimate. For other purposes, one might want to assume that the population of a given unit is distributed evenly throughout that unit so that only the proportion of the unit exposed to the LECZ would be counted. This approach is preferable to estimate the number of individuals or households that live in flood zones or who require evacuation for coastal storms, for example. When there are many adjacent units, all with partial exposure, the estimates for the partially exposed areas may then be added together to get an estimate of total exposure within the flood zone, rather than for individual administrative units. There are also more complicated rules about how to estimate population exposure based on assumptions about uneven distribution of population within spatial units. Because the answers depend on the assumptions used, it is essential to make the assumptions explicit. A common example of this type of assumption would be to use aerial photography or satellite imagery to identify built environment and density, and then to apply proportions of the population to these areas.

At present, few countries collect and report fine-grained details on census units. This is only partly because reporting is desirable for politically or administratively viable units. Another reason is because historically it has been difficult to process, manage, analyse and disseminate many more variables for many more units. However, with increasing computation power and capacity, this limitation hardly applies even in poor countries. Another concern—and one which remains quite

real is the need to preserve the confidentiality of individuals who have completed the census. As the reporting unit becomes finer – for example, down to the smallest enumeration area (EA) – the breadth of information that is used for analysis internally and reported by national statistical offices typically diminishes – in part so that individuals may not be identified through 'attribute disclosure' (REF) For the smallest units, it is common for very limited information(typically population counts, perhaps by age and sex) to be collected, and even if collected, then reported whereas for larger units data are often made available on household incomes or basic needs, race, educational and housing characteristics. There is wide variation in censuses across the globe in what variables are made available (Chamie, 2005), and even more variation in the information that is available for the smallest administrative units.

In the many industrialized countries and increasingly in newly industrializing countries, there is a good deal of census information available below the first administrative level (that is, state or provinces, typically). Data released for counties and even sub-county units – such as, census tracts, block-groups and blocks (or their equivalents) – contain more information than simple population counts (Peters and MacDonald, 2004, VIETNAM Census, 2009), but the same general principal holds: the smallest units, blocks, contain 3 variables only – population counts by age-sex groups by race. Vietnam is one such country that in recent years has not only increased the spatial resolution of its census substantially and which has collected much information below the first-order administrative units. A good deal of that information is also relevant to the climate change, which will highlight below.

Figure 1 shows province-level and Figure 2 shows province and commune-level boundaries for Vietnam overlaid with the LECZ boundary, as a close-up for one region of Vietnam. Figure 4 shows for Vietnam the province-level boundaries for the entire country where the colour hues indicate the difference in estimation of population living in the LECZ when province-level population data are used as the basis of the calculation as against sub-province (i.e., commune) level population data. At the province level, assume that the population is uniformly distributed throughout the province. Because more detailed data is available below the province level, it is known that the assumption of uniformity does not hold for population counts; it is not known, however, whether it fails to hold for other characteristics (e.g., migration rates) At least for

population counts, it can be determined what is the degree of misestimation of the population at risk that comes from a naive application of the assumption of uniformly distributed population at the province level.

The magnitude of the misestimation is shown in Table 1. In southern Vietnam, where there are entire provinces that fall fully within the LECZ, disaggregated data do not improve the estimation. But for coastal provinces, where coastal communes tend to be much more densely populated than interior communes, disaggregated data substantially affect the estimation, as indicated by the very large percentage differences noted in red. For almost all coastal provinces, using province-level data far underestimates the population at risk of coastal hazards. Four provinces are underestimated by more than 500,000 persons each. Only in one province, Hanoi, was the misestimation in the opposite direction. The province-level data result in overestimation of population at risk. Why? Hanoi city, which is densely populated, is situated at higher elevation than the surrounding areas, and above the 10 meters of the LECZ. The assumption of uniform population distribution is again false, and in this location produces an over count. Both under and over counts are problematic, particularly for agencies that might want such estimates to guide their planning. In sum, when spatially disaggregated data are available, they should be used. When they are not available, coarser-level level data may be used in this type of geographical analysis but only with caution and clear articulations of any underlying assumptions used in estimation.

The geographic size of administrative units is sometimes referred to as the intrinsic spatial resolution of census data. Unlike the resolution of grids cells, the resolution of census units is irregular. Even these smallest units are irregularly shaped and of varying sizes, as shown in Box 2. Transforming data to a grid creates compatibility with other geographic layers that are also gridded – typically physical surfaces and data that have been collected through Earth observing satellites. It is important to know the resolution of the underlying data, since it will influence the accuracy of data transformed to grids, and any additional estimates based on these grids. In particular, higher resolution of underlying data means that each grid – which can only contain a single value – will better reflect the characteristics of the area it covers.

In general, and particularly when flexibility of data usage is important, finer spatial resolution of administrative units or satellite data is considered superior to coarse-resolution data. However, higher resolution data may be more costly to process, may require greater scrutiny, and particularly when overlaying spatial data layers, the magnitude and number of mismatches between high resolution data sets are likely to be greater. In addition, for the purpose of governance and policy making, it is often necessary as well as practical to report by coarse administrative units. It is far preferable to have the ability to re-aggregate as needed in particular since some problems may cross administrative boundaries. Imagine if policy makers wanted to tally demographic characteristics for the coastal and non-coastal areas of particular provinces; fine resolution data would facilitate this though some re-aggregation would be necessary.

Although not the only question they set out to address, Lichter and colleagues (2010) recently compared three global-scale coastal zones and two population datasets to determine if there was one best dataset, or combinations of datasets, whose spatial resolution would produce the best estimates of coastal land and population. They emphasize that the datasets - and their interpretability - are very much reliant on the underlying spatial resolution and the clarity of assumptions used to produce these datasets. They find that there is no one best data set or combination of datasets, and that datasets need to be evaluated in part by their appropriateness for their intended use. They conclude with a familiar plea for transparency: "The provision of unambiguous definitions of the extent of the coastal zone, as well as of thorough and detailed descriptions of the methods and data employed and assumptions made for estimating area and coastal population, will enable the comparative evaluation of the results of different". At a local scale, sometimes much more can be said, and higher-resolution inputs of all types may be available. The recent study by Byravan, Rajan and Rangarajan (2010) on infrastructure at risk of Sea Level Rise in Tamil Nadu, India is one such example demonstrating the extent of what can be done with local data and with fewer comparability concerns (though the article is only relevant in terms of LECZ, not population) . But these examples, in more and less developed countries alike, are few and far between.

Scale of Population Data

Demographic data are increasingly available for small census units. Japan is the only country that appears to make its census available in gridded formats[2]. Yet, to date, only population counts are easily obtainable for fine-scale cross-disciplinary work. Many limitations arise from not having finely resolved demographic data. This is a particular concern for data that describe aspects of the population composition. Though only population counts have typically been available at a fine-scale, age and sex composition are usually available at that scale. However, other variables of interest that describe the vulnerability of the population or their homes (such as, education, housing, race, linguistic isolation) are not typically available at the finest scale. Statistical methods may be used with variables available at different spatial resolutions to infer attributes to a finer resolution than that which it is available, though these methods are methods are relatively new and computationally and human-resource demanding (cf., Balk et al, 2009; Elbers, Lanjouw and Lanjouw, 2003). With the use of statistical techniques, data producers and users must become more aware of underlying methodology and assumptions used to generate estimates.

Summarizing, a richer understanding of the demographic characteristics and future of climate-vulnerable areas such as coastal communities depends on the spatial information. Advances in the resolution of demographic data, the precision and agreement of coastlines with administrative boundaries, and new methods for data integration will make that possible.

While no study to date has treated a coastal region as an entity for estimating future population, with increasing seaward hazards associated with climate change in the coming decades, this is a reasonable goal to be shared by the demographic and environmental science (or coastal science) communities. National statistical offices can play a critical role in this objective because they have exclusive access to the underlying census micro data to make such estimation and forecasting possible.

---

[2] *"The unit of area subdivided by grid mesh of about 1 km square is called standard grid-mesh and shows various statistical data. Statistics Bureau, Ministry of Internal Affairs and Communications has been organizing the data of the Population Census and the Establishment and Enterprise Census into further subdivided 1/2 grid-square meshes measuring approx. 500m x 500m". See* http://www.stat.go.jp/english/data/mesh/index.htm

That is, census micro-data are not publicly released to protect the confidentiality of the population. Even samples of micro-data, when they are made available, are tend to be anonymized and can only be summarized by fairly coarse spatial units. However, census micro-data can be summarized by any geographic entity including ones that are not the administrative in nature and treated like any other geographic entity. By acquiring new spatial skills the full power of census micro-data can be used within National Statistical Offices (perhaps in collaboration with counterparts from agencies with geographic specialists).

Temporal Scale

The spatial units corresponding to census report units change over time. Country boundaries change infrequently, but sub-province boundaries change regularly. Change is expected in some areas more than others. For example, in fast-growing cities, the boundaries and intra-urban subdivisions change because the city is expanding both in population and spatial dimensions. Creating equivalencies between units over time require knowledge and documentation of the change as well as a set of decision rules on how to create equivalencies over time. Some analyst may wish to couch everything on the current set of boundaries, others may choose the older set of boundaries, and even others will want to create a gridded transformation and then letting the assumptions of the gridding process adjudicate the changes. Dealing with creating equivalencies over time between spatial variables that change is not entirely different than working with attributes that change, as commonly happens between decennial censuses. Because changes over time are intrinsic to censuses, it behooves census takers to make sure to spatial data for each point in time is maintained and documented. This will allow agency users and downstream analysts the ability to decide how to create equivalencies between units which have changed over time. Gridding sometimes offers an approach that allows for attributes belonging to different administrative units in time t and t+10 to be compared. Guidelines for managing these spatial changes are well articulated in UN Handbook on Geospatial Infrastructure in Support of Census Activities (2009).

Data Integration

Data sets on populations and data sets on climate patterns can be used together to help understand the interaction between population and

climate change. Yet no single data set provides a complete picture of individuals and the communities and environment in which they live, making a comprehensive understanding of the impact of climate change on populations difficult. Complete understanding can only be achieved by combining data from different sources, a practice that is increasing possible, but still poses many challenges. Data integration between two data sets that share identifying units can be straightforward, but data inconsistency within and between places may be non-trivial (Balk et al., 2009b).

Many examples here are given with respect to a low-elevation coastal zone, but there are many others that could be considered. For example, temperature and rainfall models (or surfaces created from observational data), aridity zones, drought scenarios, malaria endemicity zones, and flood plains are other possible climate-specific data for which one might want to construct estimates of populations at risk. The spatial data delineating each of these zones would need to co-registered with population data, so that mismatches as shown in Box 2 do not occur. That is, each data set will need to be vetted with respect to the population data, as no standard set of coast-lines whether rendered via vector or gridded format, exist. The same would apply for each additional layer, including those representing infrastructure, housing or the built environment. It cannot be assumed that boundaries for data even produced by a single national statistical office will use the same set of coastlines, water-ways and other feature that may impact estimates derived from overlays.

Box 2 shows some concerns over agreement on the precision and accuracy of data layers when more than one spatial data layer is used to generate an estimate of populations at risk. There is no consensus on how to deal with multiple data layers. The first principle to apply is one that does no harm to the estimates. A second principle is to apply spatial uncertainty. This would allow for a range for population at risk for example, by estimating exposure in areas where the grid perhaps should have been if it really should have been shifted 'upward' thereby removing the white-patches of sea that are not covered by the LECZ in Box 2.Since demographic forecasts are produced by multiple scenarios, the idea of apply spatial uncertainty should be something that is conceptually (if not technically) palatable.

Critical Steps:

1. *Identify the smallest spatial unit available from the census – i.e., the smallest for which data are available and digitized maps exist.*
2. *Identify the key indicators of interest, and the variables and crosstabs that compose them, for aggregation from micro data to small area polygon data.*
3. *Identify relevant other geographies and data: low elevation coastal zones, flood plains, temperature data, precipitation data, drylands, other types of ecosystems, etc.*
4. *Based on the criteria in the chapter above, decide whether the analysis will use gridded population data in concert with raster environment/climate data, or will use polygons or zonal statistics derived from raster environment/climate data in conjunction with the existing small area geography. Conduct the relevant transformations.*
5. *Identify and attempt to correct sources of error in the use of data from multiple sources. These can include geographic variations like different coastlines, as discussed above. They can also include small area polygons from the census that deviate from social boundaries; overlaying of small area boundaries on aerial photography or satellite images can help in this exercise, and spatial software provides tools to adjust census geography to better match what is found on the ground.*