Towards a general theory of measurement

Essay by Nick Chrisman, 2018-12-17.

Preamble

This document has a (silly) over-general title, but eventually we need a General Theory of Measurement, or at least workable Models of Measurement to apply to cases like the current exercise at INEGI (four letter acronym escapes me right now).

The background for this essay comes from decades of experience in interdisciplinary communication. Kuhn's 'paradigms' have many interpretations (not the least his own various attempts to explain them). For me, the key moment for each discipline comes when deciding to measure something. First they have to agree what the 'somethings' are… Then they come up with the framework to ensure that the measurement can be repeated – or at least usable by others.

In the UN-HLG-MOS schema, there are two preliminary steps (Design, Infrastructure) that depend on creating (or assuming) a model for measurement. In my experience, the geospatial sector uses more diverse measurement frameworks than most other disciplines. To some extent our toolkit is all about changing the measurement framework (not just living inside one of them). Perhaps that is due to the historical nature of the different arts and crafts that emerging in Geographic Information Science, perhaps it is due to the diverse expectations placed on the field by its users (and patrons). But the origins are less important than setting out a broad scheme that can encompass the activities across a big agency like INEGI.

Example: a simple difference of perspective

Let's start with a simple case to explain how this essay will proceed. Medical doctors (let's call them MDs) have a diagnostic category 'cancer' (yes, many many sub-cases, but we will work at the general level). It applies to their 'unit of analysis' – the patient. In fact, you don't really have to do a lot of specification, MDs operate at the level of 'patient' in their practice just about all the time. (Yes, infectious disease experts may introduce some aggregations like household, neighborhood, etc. but that is not the average work of a clinician.) So, the clinical view: patients can 'have' cancer.

From a biological research perspective, the organism is not the basic unit of analysis. A person is made up of cells. Cells are the unit that can 'have' cancer. If enough cells have cancer, the organism (patient) will show signs that the MD can diagnose.

Thus, the concept 'cancer' is operationalized differently by different disciplines (Joan Fujimura and colleagues wrote a nice piece on this topic, Fujimura, et al. 1992; Harvey & Chrisman 1998). It may look like some simple aggregation rule: if patient has greater than k cells with cancer, then patient has cancer. Of course, an organism has a whole range of processes that operate BETWEEN the cells, and the operational diagnosis may be when the immune system has failed to kill off/ restrain the cancer cells. But this is not an essay about cancer.

Statistical frameworks- an audacious glittering generalization

Many disciplines use various forms of statistical methods, often adopting statistical procedures that may or may not fit their needs. At the base, the generic statistical model has some 'unit of analysis' – which we will call a ‘case’, since that is the usage in many fields from medicine to social science. Cases are assumed to be sampled from some broader ‘population’. The sampling introduces a potential for error (in that the particular cases available may lead to a difference in the measurement obtained). So, handbook stats package model (we could call this Fisher’s paradigm): big (unknowable) population, particular (smaller) sample. Models are needed to understand how a measure obtained from the sample (even straight-forward measures like the mean of some numerical score or physical measure) might provide sufficient knowledge about the population. The models deal with stochastic (chance) elements. The science of statistics is highly developed on various stochastic models. Basic direction of logic: induction, from specific to some (nebulous) generality.

The peculiar thing is that modern statistics developed from the experience of ‘states’ measuring information of use in statecraft (what states do: administration above all). In particular, the modern nation state emerged in the late eighteenth century/early nineteenth century with some general principles of universal application. For example, everybody had to pay taxes (a generalization that always has some loopholes). The laws and regulations apply in every part of the territory. The elementary tools in this original form of state-knowledge were registers and a census. In principle, the census measures the whole ‘population’ in a very precise meaning (not that Fisherian generality of all the possible worlds). Of course, any real-world census falls short; it cannot capture everything, but it certainly sets out to plug any leakage. It is rather harder to assert that the actual results of a census are a nicely random sample from the larger population. The deviations are likely to be rather systematic, and badly distributed. (Depending on method, but homeless people are much less likely to be captured in the US Census these days. Extraordinary measures are needed to try to obtain a reasonably complete coverage of this sub-population. These kinds of errors lead to different designs of a census. The British colonial census in India was not based on housing units, but enumerating people during the middle of the day as they moved around.) What a census shares with the Fisher abstract model is a firm basis with a unit of analysis. A Census of Population counts ‘people’; a Census of Business – well it deals with enterprises, but the definition of a single ‘business’ gets a bit more abstract… Usually something is decided and it becomes the operational definition of the unit of analysis.

Unlike the Fisher attachment to testing hypotheses, the state-craft kind of statistics is designed to roll the individual cases into reporting units at higher levels of aggregation. The population is counted ‘by’ municipality or any other level, up to the entire universe of the Census (typically the nation). Since the basic units are considered unitary objects, addition is the aggregation rule; it is axiomatic, and so obvious it may never be said explicitly. Of course, any other measure (such as economic activity) can measured with some weighted value (like income) and then also added up to get accounts for aggregated reporting units.

The hierarchy of reporting units is also an element of state-craft, and thus a part of statistics without too much deliberation. Of course, reporting units are mutually exclusive and exhaustive of the universe of discourse (the nation). In practice, the set of reporting units gets complicated. Real nations rarely have a single rigid hierarchy despite official allegiance to the abstract principle.

A Census is most useful if it is repeated, since change over time is often the biggest interest in Census users. Therefore, the aggregated units (typically politically defined) only remain comparable if they are maintained unchanged between census dates. It becomes a complicated job for geospatial processing to figure out the correspondence between aggregated units after a change. There is a rather detailed literature on what is called ‘areal interpolation’. This issue will be used an example below.

Geographic measurement frameworks

Geographic measurement is my specialty, so I should stick to what I really understand.

Introducing the concept: an example

David Sinton presented a scheme to understand geographic measurement (1977), based on three roles to be assigned to the basic dimensionality of geographic measurement: time, space, and ‘attribute’. (Yes, attributes vary, but let’s explain the basic concept before we start pulling it apart.) In Sinton’s terminology, in order to measure one of the three, we have to restrain the other two as ‘fixed’ and ‘controlled’. For most maps, the time is static – fixed. Conceptually it is a snapshot in time, but that may mean a ‘period’ of validity (of variable length depending on the rate of change in the attributes in question). The role of ‘control’ is more complicated to understand. An example would be a contour line; it represents the location of a particular height value. Typically, we set an interval and we create contours for every 10m or 100m or whatever. In Florida, they use 3 inches (0.1m) for contours at 1:24,000. The interval is a matter of choice and practical considerations. What is important is that there is an integer simplification of the elevation surface. Once the height is set, the 2D location (space on the map) is free to vary. We **measure** the location at an established (**controlled**) height, on a **fixed** date/time. Contour lines have been used to represent a surface for a bit less than 200 years.

An alternative approach is to control the location, so that the height of the surface is measured in a regular 2D lattice (a grid, often rectangular). This inverts the situation of the contour line. The location is now in an integer space, while the height can freely vary. The heights can be represented by real numbers, for example (typically floating-point approximations are adequate). The lattice is typically regular on some planar projection system, though some cartographic agencies pretend that arc-seconds can be used to define a lattice (it has spherical geometry, but that is not usually an issue for the scale of an individual map). A grid as spatial framework can be described by these parameters: an initial point, an azimuth (angle of the rows), and spacing (along the row and between the rows - often the same). There is no need to represent the coordinates of each point; it is implicit in the relative location in the array. This kind of terrain representation has been called a ‘Digital Elevation Model’, but it would be more precise to call it a Digital Elevation Matrix. It is also called a Digital Terrain Model/Matrix, but that may also include some filtering related to human constructions. That is a topic for greater detail, but some terrain data tries to remove buildings (it may make sense for things like following the path of water over the ground, since buildings will do something else to water flow…).

Measurement frameworks

In my textbook and other publications, I use the term ‘measurement framework’ (Chrisman, 1997, 2001, 2016) to refer to this package of decisions. In the UN-HLG-MOS scheme, these are design decisions, I think. They are backed up by the supporting elements that the UN system calls ‘infrastructure’, since that is an over-used term in the geospatial sector, I will use distinct words for each kind of support.

The list of measurement frameworks that I have developed provide a general guide, but the scheme is over-ended. Other combinations and nuances do occur. (I have not proved transitive closure, in other words.) Let’s review the process used by INEGI for generating topographic products. The process starts with images. INEGI stores images in a registry, but it would be better to consider this an archive. Each new image gets a new identifier, and the list is open-ended. An image is a ‘raster’ data structure, using a grid-style spatial control. The space in question is on the image plane, basically each pixel is one unit wide. There is not connection to a spatial reference system. The connection is made with certain points whose ground coordinates are known, and also located in the image. A single image can be referenced to the ground, but only in a 2D sense. Like any optical system, you need stereo vision to obtain ‘depth’ – the connection to 3D. Every image point is found on two images (at least), and a least square solution solves for the location of the sensor at the time the image was taken. Photogrammetry has developed a constraint model called block adjustment that essentially solves for the speed of the airplane (satellite) and variations in yaw, pitch and roll between images. Every spot of color on the image (pixel) is projected down onto the ground surface. It ceases to be a grid structure, since space is no longer the control. It is called a ‘point cloud’. Like other objects (in what is called the vector model) the spatial position is measured (and free to vary). The image data is copied along. Point clouds have so little structure that they break the rules a bit and carry measurement of the image color and the position. However, with no inherent structure, this data structure is hard to use.

For most practical purposes, a point cloud is simply a way station to get what you need from the original image source. Most usually, the first step is to bring some spatial control back in force. A Digital Terrain Matrix (under lots of synonyms) can be generated by interpolating the best fit in a regular grid pattern from the heights in the point cloud. There are lots of splines or distance decay algorithms applied – the underlying process is to find the relevant nearest points in the point cloud (some spatial search) then generating a best fit of the value at the grid intersection point. In easy terrain, the original grid pattern of the image can make this process rather simple, but in rough terrain, there can be some grid cells with many points and other areas with sparse distributions. Most interpolations end up smoothing the surface somewhat, unless the simple nearest neighbor algorithm is applied. This just copies the value from the nearest point. If the point cloud is denser than the target grid, it works fine. In the other direction it is problematic (creating artificial flat areas).

The Digital Terrain Matrix can be used to rectify the original image, too. This time the image data is copied into a spatially referenced grid. This product is called an ‘orthoimage’, and serves as a base for certain maps products (for many environmental analyses, you can detect features by visual interpretation of the orthoimage). Automated feature detection uses various procedures to trace ‘objects’ in the image. Current algorithms tend to have origins in the robotics literature, but lots of procedures can be applied. Pixels are grouped together to detect points lines and areas, in a different ‘vector’ measurement framework. Here the identity (classification) of the object is the control, and the position is measured (to the resolution of the grid of the orthoimage).

Also, starting with the Elevation Matrix, another elevation product can be generated by estimating the location of a slice horizontal slice across the terrain. The intersection of a plane with the surface generates contour lines. The height is controlled (and usually performed in a regular spacing), the position is measured (and interpolated to any degree of resolution. Contours will still depend on the original spacing of the matrix for their ground validity. Again, the position of the contour line is estimated from nearest neighbors. The simplest method applies a lattice of triangles, but the output looks blocky unless smoothed with a spline along the contour.

Another product from the DEM fits a local surface (to obtain the slope and thus the surface normal direction). In standard practice, pixels are bright if pointing to the Northwest, and darker in cyclic logic rotating around to Southeast. The exact angle is the subject of some recent studies of map readers (Biland & Coltekin, 2017).

Thus, the photogrammetry procedures can be understood as a sequence of changing the measurement framework.

This example introduces the basic division between spatial control and the 'attribute' control of the vector model. There are some additional kinds of vector model more attuned to the Census requirements.

Many spatial objects are not visible on an image source. For example, the boundaries of states, municipalities and other political/administrative zones need to be generated from some source. Sometimes there are direct surveys performed for important objects like international boundaries. If the spatial measurements are provided, cartographers generate a polygon object directly. Space is measured, for the controlled nomenclature. In this case, there should be a fixed list of objects (States of the Republic, etc.) These codes are the control for the measurement of the position.

The attribute table for each polygon can be loaded up by joining any other table whose primary key is the coded name of the entity. For example, median income of households can be attached. This is a continuous value, attached to a categorical entity. A thematic map of household income will show a sharp edge at the boundary of the polygon. This is what cartographers call a 'choropleth map'. They spend a lot of time designing the class intervals and other details of the display. The sometime think that the income data was 'measured' in some way. The user needs to remember that this is a summary over the whole spatial entity. The boundaries are not the boundaries of income data, just the administrative boundaries. In my scheme, choropleth maps are composites, with geocodes as the primary control for the geometry of the boundaries, then a seemingly continuous measure poured into the polygons. The stairstep surfaces can be visualized as raised prisms, though the underlying distribution is likely smoother.

As support for choropleth thematic maps, the codes for the various spatial entities in the Census tables have to be used in the control of the corresponding maps. A 'registry' of these defined codes is the support required. There should be a one-to-one relation between a code and its polygon (with the additional concern for objects with multiple 'exclaves' or 'islands'). The collection of polygons coded for an object need to be considered as a single entity in the attribute table (for joins). A completeness check will ensure that all the codes in the list are represented once.

By contrast, other polygon objects have a different origin. Categorical categories of a land cover classification may be recognized (by whatever means: photointerpretation, digital classification, clustering, maximum likelihood, Bayesian whatever…) and a boundary drawn to distinguish it from the neighboring category. The measurement framework is a simpler exhaustive coverage that I call 'categorical coverage' (my friends at ESRI may have used this to name all vector layers as 'coverages', but named objects like states or muncipios are different from these categories. Here each polygon classified 'urban' does not attach as an exclave of the other urban zones. Each object is just another case of 'urban'. Eventually they can be given distinct names, perhaps, but the polygons of pasture, rainforest, bosque… will defy easy naming. They are just instances of the category. One-to-many in a database sense. All codes do not have to be used. In some international land cover system, there may be a code for tundra, which is not present in Chiapas, and 'selva' is not present in Sonora. In some cases, the adjacency of classes can provide a consistency check. A wetland (zonas humides) might not be surrounded by desert, without some intermediary class (like grass or trees). The registry might want to have a standard set of codes of land use or land cover, but often in the case of unsupervised cluster detection, the categories may not align with any standard.

Areal interpolation

Dependence on measurement frameworks can be best seen when changing levels of aggregation. There are some measures that are easy to aggregate and others that are easier to disaggregate. Let's use some concrete examples. Population COUNTS obtained for some unit area can simply be added together to create a population count for a larger region – as long as the enclosing region includes the whole original area. If the original unit is split, it may be difficult to allocate the true count for the subareas without significant processing or additional information. There is a specialized field of geographic analysis, dating back to the 1930s, called dasymetric mapping. It brings in ancillary coverages (like land use) to estimate density/population inside collection units.

Going the other way, some attributes (like population counts) cannot be disaggregated without additional information. Other attributes, of a categorical kind, particularly, are easier to disaggregate. Inside a large urban region, all subunits are also urban. These rules of aggregation are not written down in most cases. They are really a sign of how much processing is required and what assumptions are needed.

Scale

Geographic information has traditionally been linked to a complex metric: 'scale'. The term comes from the representative fraction of a paper map. Distances on the ground are reduced to fit on the map sheet. However, even on the scale of 1:1, a map would still have to make decisions that will reduce any sense of direct correspondence (despite the fanciful stories of Jonathon Swift and Garcia Lorca). A map or a geographic database have to decide what entities to recognize and how to represent them. This is the core of a measurement framework. Are streets represented as a line or an area (two-sided polygon between block faces)? At what level of resolution do certain object disappear from the map? Do any objects take priority over their neighbors? (Often water features (shorelines, rivers) are placed first; railroads are placed next, then roads and other linear objects have to adjust.) If the database is derived from the map, the features may have been filtered and displaced.

For INEGI, the basic database may need to record the actual position as best as possible, leaving any scale changing and displacement for a derived product at a given scale. If this has not been the case in the past, it will take some effort to make the transition.

Specific issues for INEGI

The relationship of mapping products and Census operations create a number of practical questions that may take time to resolve. The Registro de Careteras comes from an external source. INEGI may need to understand issues of completeness and attribute accuracy (coding) of this data source. If additional streets are detected in field work or from imagery, how do these objects integrate with the registry? Items in the registry can be used to conduct completeness/ logical consistency checks between the road registry and the spatial data. Individual block faces should aggregate up to a 'street' object if they chain together in some linear fashion, but two parallel block faces should not be coded from a single road item in the registry (Maybe exceptions exist?).

It was stated that 30% of households do not have an address. Considering the massive effort in conducting the Census (Economic, Population, Agro-pecario), it should be a priority to determine how the concept of addresses should be extended, or revisions provided to get closer to complete coverage. It may not be possible for homeless people, or in very remote rural areas.

Beyond the simple ‘object’ measurement

One additional measurement framework applies to flow data, such as trade accounts (inter-regional as well as international). Here, the control is the pair of objects. The flow between the pair is measured. Logically, a trade database is a matrix of objects times objects. It is standard practice in regional input-output analysis. Cartographers have spent some effort on means to display flow data, in a planar manner or with 3D effects. GIS software packages are sorely under-equipped to store and display this measurement framework.

Measurement frameworks for statistical data

The terminology developed here comes from the circumstances of geospatial data, but it could also be applied to measurement through surveys and administrative systems. The concept of ‘control’ and ‘measurement’ may be considered obvious in cases where there is little controversy about procedures. But a Census that works by household connects to individual persons indirectly. Rules are needed to ensure that person is not counted twice, for example. Businesses are even more slippery; one operation can have locations spread across a city (or a country). What is the level of aggregation for this census?

Agriculture has a link to the land, so parcels and fields are objects that make a difference. The cadaster provides a set of exhaustive objects in some regions to ensure that land is not counted twice. In addition, land cover data can provide a cross-check that portions of a parcel might not be in full use. Area measurements from the mapping source can be used to cross-check the reports from respondents (perhaps requiring adjustment for area on the terrain, rather than planimetric areas).

Summary

The concept of a ‘measurement framework’ provides a useful means to organize the metadata required for different cases. While it was developed in the geospatial sector, particularly to distinguish the difference between raster and vector approaches, it can also apply to items in the statistical sector that use different sample frames. What it offers is an explanation of which items belong in a ‘register’ and which are simply ‘archives’ of accumulated sources.

References:

Biland, J. & Çöltekin, A. 2017: *An empirical assessment of the impact of the light direction on the relief inversion effect in shaded relief maps: NNW is better than NW,* **Cartography and Geographic Information Science,** 44:4, 358-372, DOI: 10.1080/15230406.2016.1185647

Chrisman, N.R., 1997: **Exploring Geographic Information Systems.** New York: Wiley. 298pp. (First edition).

Chrisman, N.R., 2001: **Exploring Geographic Information Systems.** New York: Wiley. 270 pp. (Second edition) (794 citations to the two editions)

Chrisman, N. R., 2017: *Measurement Theory.* in **The International Encyclopedia of Geography**, edited by D. Richardson N. Castree, M. Goodchild, A. Kobayashi and R. Marston. Hoboken: Wiley. DOI: 10.1002/9781118786352.wbieg0438

Eicher, C. & Brewer, C. (2001) Dasymetric Mapping and Areal Interpolation: Implementation and Evaluation, Cartography and Geographic Information Science, 28:2, 125-138, DOI: 10.1559/152304001782173727

Fujimura J H, 1992: *Crafting science: standardized packages, boundary objects, and 'translation',* in *Science as Practice and Culture* Ed. A Pickering (University of Chicago Press, Chicago, IL) 168-211.

Harvey, Francis and Chrisman, N. 1998: *Boundary objects and the social construction of GIS technology.* **Environment and Planning A,** 30(9): 1683-1694. (288 citations)

Zandbergen, P. & Ignizio, D. A. (2010) Comparison of Dasymetric Mapping Techniques for Small-Area Population Estimates, Cartography and Geographic Information Science, 37:3, 199-214, DOI: 10.1559/152304010792194985