National Pandemic GIS & Informatics Task Force Playbook

December 2022

Developed in collaboration with URISA and NSGIC

Produced by: NAPSG Foundation

Table of Contents

Introduction

During the COVID-19 Pandemic, tuning into the evening news meant seeing a Geographic Information System (GIS) display that showed the spread of the disease. Yet, the geospatial industry nationwide knew it had more to offer in addressing the pandemic.

It is clear to GIS professionals that the spread, impact, and effectiveness of response efforts are best understood in contexts of space and time. With that understanding, this playbook attempts to address pandemic policy and response. Policies, actions, and communications work best when informed by data.

A key benefit of spatial data is that data sources can be combined based simply on location. These combined geographic data can offer an understanding of a geographic area and explain variations in the metrics about a pandemic. These insights also can inform action. Spatial data, for example, can help explain why one neighborhood is experiencing a higher spread-rate over another, or data can predict the expected hospitalization or mortality rate based on the prevalence of comorbidities in the area. A key to unlocking the value of existing spatial data like demographics, health characteristics, or even levels of trust in government is managing and publishing pandemic data using the recommendations in this playbook. Sound data practices and protocols can greatly impact the quality and timeliness of decisions. Data about the pandemic, such as the spread rate or vaccination rate, can be processed and formatted differently for different audiences. Statisticians and epidemiologists might need the data in a raw and unfiltered machine-readable form. Emergency response staff might benefit from an action-focused map dashboard, while policy makers might need a different dashboard to help monitor critical disease and staffing metrics.

However, the ultimate decision makers are the public. Based on the messaging, levels of trust, and level of engagement with government, the public will decide where to comply with control measures or simply ignore them. One of the keys to building public trust is in the transparency and quality of the data published. Consistently maintained sets of pandemic data can satisfy the data needs of all these audiences when managed, processed, and published appropriately.

This Playbook is not intended to be a cover to cover read. The first section lays out common scenarios, stages, or business problems, while the second section provides tactics that apply to one or more of the actions covered in the first section.

Scenarios, stages, or business problems

Tracking Cases

From a GIS perspective, tracking cases is about five things:

1. Collecting and storing valid home address locations of those tested.
2. Geocoding those addresses to produce geographic coordinates.
3. Deduplicating records or eliminating duplicate copies of repeating data.
4. Aggregating those results to an appropriate unit of geography (area) to provide relevant insights.
5. Publishing the case data and rates.

Locating Home Addresses

Home addresses are critical if the data is to be analyzed alongside demographic data from sources like the US Census Bureau. It is important to understand that location addresses (i.e., where people live) are not always the same as mailing addresses. When addresses are collected, they should be the address one would use when ordering a pizza or making a 911 call. An easy way to remember this is that you would not have a pizza delivered to your P.O. Box.

Geocoding Addresses

An address geocoder is simply a URL ready to accept an address and return a correctly spelled and formatted address along with the x and y coordinates associated with the address. Contact the GIS professionals in your area for advice and assistance. See *GIS Tactics, Address geocoding/location address validation*.

Deduplicating Records

Once geographic coordinates and correctly formatted addresses are associated with individual test records, duplicates can be found using a combination of address, name, and date of birth. While the duplicate records are useful for measuring how many tests were performed, they should be appropriately addressed before establishing an infection rate.

Processes to rapidly transfer records that fall out of jurisdiction and identify and merge persons with multiple profiles should be a well-resourced effort throughout a response to ensure data quality. Automated processes using fuzzy logic (e.g., single data source demographic matches) to identify potential duplicates should not be solely relied upon to address data quality issues. Pre-pandemic efforts to establish master person index using multiple data sources offers greater opportunity to successfully employ automated deduplication processes.

Communications teams should be briefed on how to handle questions around negative daily case counts. These often occur when jurisdictions report net daily counts. During disease surges, local health departments are not able to conduct deduplication efforts on all cases. As a result, between waves, data quality efforts are increased and can result in a negative daily case count.

Aggregating to a Geographic Area

Once geographic coordinates are associated with individual test records, those records can be automatically aggregated to any unit of geography for which GIS data exists. The most common units of geography for case data aggregation are country, state, county, and census tract. There is no need to ask people being tested for any of these data elements as they can all be automatically generated from the coordinates of the address using GIS tools. For understanding a world-wide or nation-wide pattern of spread, county and state-level aggregation is appropriate.

County-level aggregation is appropriate for understanding a general pattern within a state or for communicating restrictions on traveler business operations. However, even county-level aggregation of case data does not provide the spatial resolution adequate to understand the localized factors impacting the spread at a neighborhood level. Both spread rates and the causes of the spread vary widely from neighborhood to neighborhood. Aggregating case data to census tracts is ideal for understanding local variances in the spread while anonymizing the data to protect privacy. Census tract level is adequate to inform policy and take action in areas of concern. A common mistake is to aggregate data to zip codes. Zip codes are created purely for the efficient distribution of mail and do not cleanly overlay any other units of geography. See [GIS - Tactics Spatial Data Aggregation to Anonymize Data].

Spotlight: **Zip Codes**

For a better understanding of why you should not use zip codes for geospatial analysis, click here.

Publishing the Case Data and Rates

Once case data is aggregated, the resulting count or rate data is no longer sensitive and can be shared publicly. Public data allows for broad collaboration without the cost and burden of private data transfer and user account management. Public data also allows researchers, statisticians, and external epidemiologists to provide insights, while also making aggregation to larger units of geography more reliable and efficient. Although the concern may exist that some will misuse the public data to undermine policy decisions, public data and the transparency of sound science can have the effect of increasing public confidence in government and form the foundation for consistent public messaging.

Prior to publicly publishing any aggregated health information, HIPPA compliance officers should be consulted. Guidance and accepted practices are not universal across state and local jurisdictions and may change during a pandemic. Tactics such as masking low case count areas and combining areas with low populations or case counts should be considered. Processes for publicly released data should be designed to remain consistent throughout a pandemic. For example, when the case incidence rate is low, switching from daily reporting to weekly should be well communicated. A daily dashboard reflecting data that changes weekly could inappropriately show spikes on the day of update. See *GIS Tactics, Spatial Resolution*.

Spotlight: Louisiana Case Data by Census Tract

The State of Louisiana published case data by census tract that did not receive negative responses or complaints. Before publishing said data, Louisiana removed some false anomalies, which in some cases the entities used a default address for anyone who refused to provide an address or any unusable addresses. This resulted in inflated numbers in the census tracts containing the default addresses. Louisiana only included records that geocoded to an address point or street centerline. Some geocoders will place a point on the center of a zip code if that is all that can be found or matched. Records geocoded to a zip code centroid are often located in the wrong census tract.

Understanding the Sampled Population & Disease Transmission

Water Waste Surveillance

Testing wastewater for viral nucleic acid (i.e., severe acute respiratory syndrome coronavirus 2 RNA) is an approach for gaining an early indication of presence and relative disease burden in a geographic area contributing to the sewage systems. Along with the target pathogen (e.g., SARS-CoV-2), a second pathogen that is normally shed by everyone can be used as a control. The geospatial data contribution to wastewater surveillance is the ability to understand the geographic extent of the population contributing to the sample. Property parcels with associated property assessment records can be used to determine those parcels upstream of the sample and connected to public sewer (as opposed to private septic). If a sample is taken at a sewage treatment plant, the geographic extent is represented by all the parcels in the sewer shed.

Temporal considerations should be given to the time-of-day samples are taken, since samples make their way to the sewage treatment plant as workers make their way to the office or children to their school. The lifestyle patterns impacting when people flush a sample should be considered.

Samples can also be taken at specific maintenance hole locations upstream of the sewage treatment plant to narrow in on the location of a localized outbreak. In this case, GIS data on the layout of the sewer network along with parcel data can be used to delineate the geography represented by the samples. By using parcel data, assessment records can differentiate private septic systems from public sewer connections

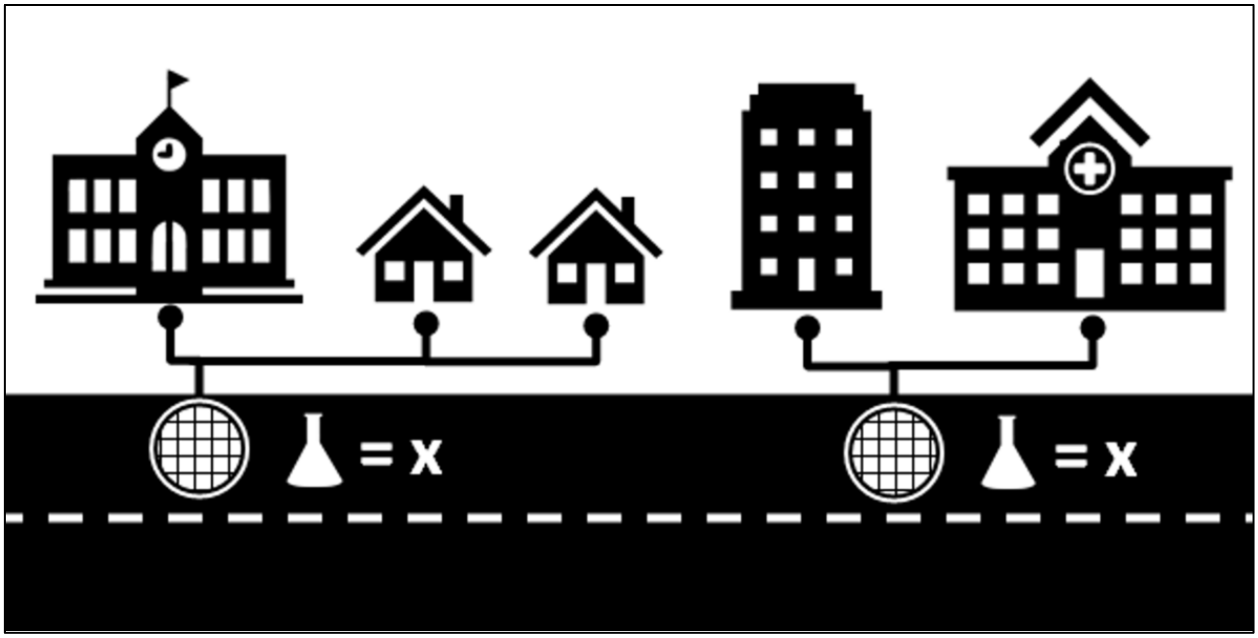

Figure 1: An example of how RHI data is provided based on where the sewage sample is taken.

Spotlight: Using Wastewater Surveillance Data to Support COVID-19 Response

The Center for Disease Control reported on Utah's wastewater surveillance data, stating the data had been used to help direct clinical testing resources to areas experiencing increased SARS-CoV-2 RNA levels in wastewater. Beginning in 2021, wastewater data are one of the main components of a ranking system to determine where to send mobile testing teams. Wastewater data have also been used when interpreting clinical case surveillance data.

Understanding Disease Burden

During the COVID-19 Pandemic, disease incidence rates varied widely across the country, and these variations were often very localized. In New York City, for example, neighborhoods on the Upper East Side experienced relatively low infection rates, while neighborhoods just a stone's throw away experienced much higher rates. Many workers living in the Upper East Side maintained employment suitable for remote work, while workers in nearby neighborhoods may have had a choice between feeding their families and social distancing. In these neighborhoods, most people work in transit, retail, or in other in-person service jobs.

Spatial regression techniques can provide a statistical understanding of the localized variances in disease burden. These techniques can quantify the percentage of variance in the distribution explained by data representing distinctive characteristics of the neighborhood. Spatial regression can also provide coefficients on each of these characteristics, creating a mathematical equation predicting the distribution in an un-effected area with similar characteristics.

Variable	% Variance explained	Coefficient
Household size	A	V
Income	B	W
Ethnicity	C	X
Cell phone tracks mobility index *	D	Y
Cell phone tracks from tract to tract**	E	Z

Table 1: Examples of variables that could be used in spatial regression.

Disease incidence rates aggregated by census tract can be analyzed against any existing data aggregated to the same unit of geography. Important data sources for the independent variables used in spatial regression can include census demographic data, aggregated mobility (cell phone track) data and more. Importantly, standard methods are needed to aggregate data to static units of geography (in this case census tracts) for the previously mentioned statistics to be valid.

***Cell phone tracks mobility index**
Commercial aggregators of cell phone movement data can provide an index qualifying the relative measure of movement within a census tract on any temporal cycle. These indices can be normalized by daytime or nighttime population for a per capita movement. Note that census population counts reflect where people sleep, LandScan (Oak Ridge National Lab) population estimates reflect "ambient population" (average over 24 hours). Commercial sources of business locations and daytime population are also available.

****Cell phone tracks from tract to tract**
Commercial aggregators of cell phone movement data can also provide an index qualifying the movements from one area to another.

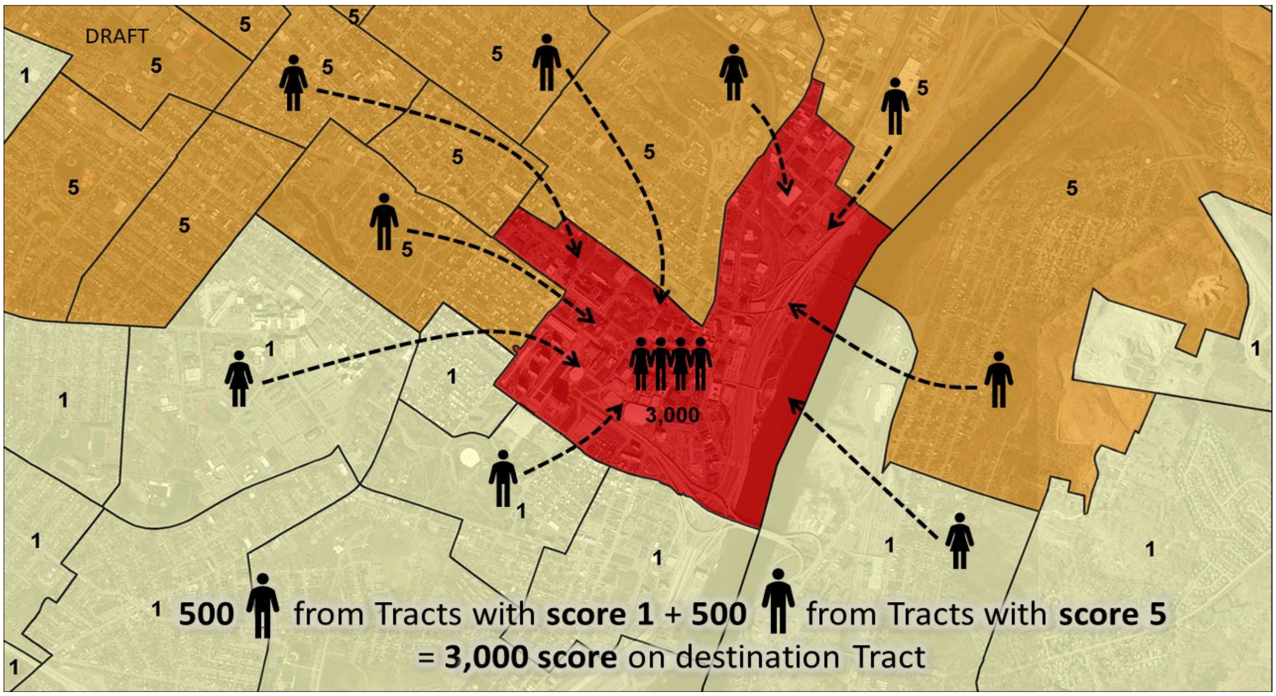

Figure 2: In this example, each census tract is given an infection score based on the pandemic infection rate of that tract.

In Figure 2, scores of 1 or 5 indicate the relative infections rate. Every cell phone moving into the tract being analyzed brings with it the score from the tract it came from.

500 phones came from tracts with a score of 1 =	500
500 phones came from tracts with a score of 5 =	2,500
Total inbound score =	3,000

Repeating this procedure for every tract would provide a relative inbound score for each census tract, thereby creating another variable in the spatial regression procedure.

These scores can be tuned using criteria such as how long the cell phone stopped in the census tract, ensuring that anyone driving through a census tract is not counted. Using this method, cell movement data modeling can be kept simple or can be refined to model specific use cases (e.g., possible infection spread on mass transit).

The spatial regression techniques create a coefficient to each variable in Table 1, providing a predictive formula of spread rates:

	V x Household size
+	W x Income
+	X x Ethnicity
+	Y x Cell phone tracks mobility index
+	Z x Cell phone tracks from tract to tract
=	Predicted spread rates

Spotlight: Controlling spread in Nursing Homes

Dr. Keith Chen, from UCLA, used mobility data along with location data to depict the nation's nursing homes in order to create measure of how connected nursing homes are to one another through a network of shared staff. Dr. Chen's assertion is that 200,000 lives could have been saved if policies were put in place allowing nursing home staff to pick just one facility to work in.

Tracking Vaccination

Placing and administrating vaccination sites

Mass vaccination of school-aged persons using a school-based approach offers many opportunities for directing limited resources and monitoring progress. Schools maintain high quality records on their student populations, including aggregated race and ethnicity counts. In many jurisdictions, students must attend a school in a geographically defined catchment area. Geocoding vaccination data from Immunization Information Systems (IIS) for school-aged persons and aggregating by school catchment area allows the numerator to be estimated for the proportion of students vaccinated by school. Jurisdictions can then use this information to organize school-based vaccination clinics with parents or using in loco parentis. Spatial data needed to make sense of school-based vaccination rates are school catchment areas. Establishing the connections between the schools' GIS data and the vaccination data collection efforts can happen in the pre-pandemic planning phase.

Collecting Vaccination Location Data

Aggregating to a geographic area

Just as geocoding and aggregating individual pandemic test results provides insight into the spread of a pandemic, geocoding and aggregating individual vaccination records provides insights into the success and shortcoming of a localities mass vaccination efforts. Geocoding individual vaccination records results in thousands or millions of dots, each dot represents something that could be a HIPPA concern. Aggregating those dots to a unit of geography like census tract anonymizes the data to make it no longer about the individual but rather the characteristics of a population in a geographic area. This anonymized data becomes more useful for many policy and strategy decisions than the individual records.

See *Address geocoding/location address validation*. The same techniques described above in the *Understanding the Spread* can be applied to aggregated vaccination data to understand and remedy variations in vaccination rates.

Aggregation, if in sufficient numbers, allows for public use of otherwise protected health information. Also, when data such as case incidence rates are standardized, it allows for comparisons to be made across differently sized or populated geographic areas. Insights such as languages spoken and vehicle ownership have proven important in vaccination planning and implementation. A key strategy is to publish vaccination data anonymized to the census polygons (i.e., tract, group, county) of the finest resolution possible while considering personal privacy. In many cases, census tract is the sweet spot. The data supply chain is critical to consider and communicate. Acquiring a valid location address and validating it against a geocoder needs to be upfront in the data collection process. In some cases, this will happen in a vaccination appointment system. In others, it will happen at the vaccination site. The systems used to capture addresses can be linked automatically to a standard geocoder Application Programming Interface (API) without adding a burden to vaccination providers.

In the case of nation-wide pharmacy chains capturing vaccination data, a nation-wide standard geocoder would be ideal. More than half of the states are contributing their address points to the National Address Database, which in the future will be the data source of a nationwide geocoder API.

GIS Tactics

Address Geocoding/Location Address Validation

When preparing to geocode addresses, the first thing to understand is the difference between postal addresses and location addresses. Location address and mailing address can be the same, but they often are not. For example, there is no local mail delivery in the village of Altamont, NY. Everyone who lives there picks up their mail at the post office. However, each residence has an address for 911 dispatch. Be sure to focus on location addresses, and do not allow P.O. Boxes.

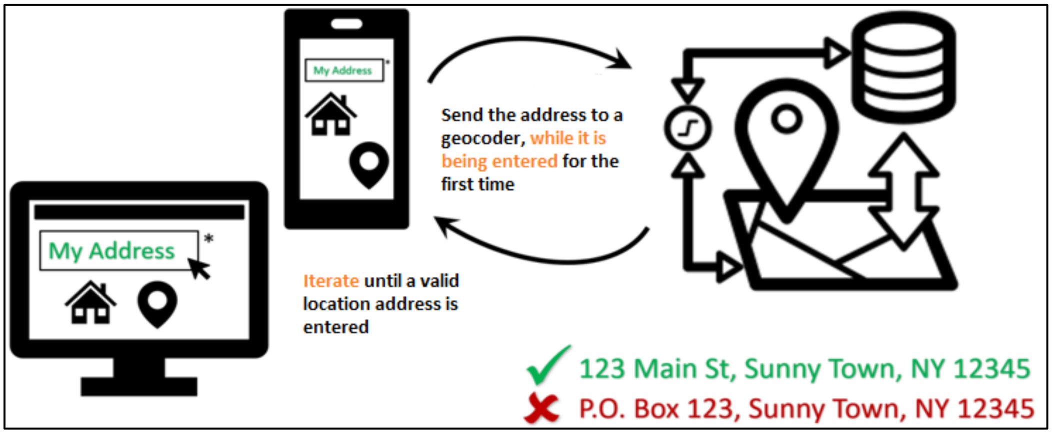

Figure 4: Explains the process of finding a valid location address by a geocoder

The best time to validate addresses is during input. Whether in a call center or on an on-line form, the address input box should be connected to an address geocoder. The developer of the entry form can embed the URL for the geocoder right in the form. Without this, once the caller hangs up the phone or hits submit, the opportunity to get the address validated, may be gone. For example, if someone enters 400 Pearl St. Albany, NY, the entry form would need to present the two choices of 400 North Pearl St or 400 South Pearl Street. Those two points are over two miles apart.

A **Geocoding API**. This will:

- **validate the address** – providing corrections
- return **geographic coordinates**
- **geographic coordinates can be stored with the valid address**

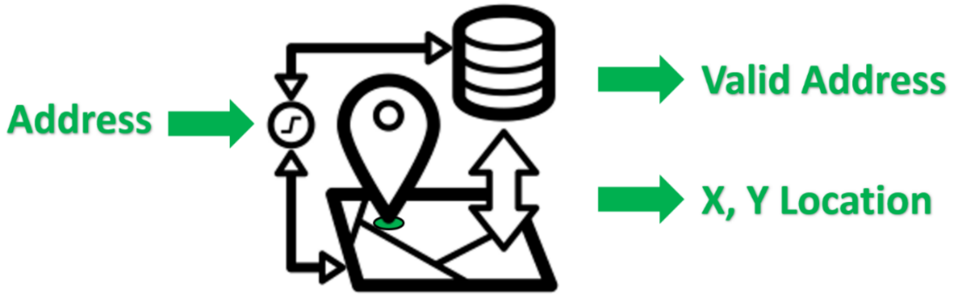

Figure 3: Step 1 is to geocode addresses into geographic coordinates.

Spatial Data Aggregation to Anonymize Data

Once you have coordinates on addresses, records can be aggregated to any unit of geography. By combining multiple individual records into a sufficiently sized count or rate, the data is anonymized and no longer sensitive as it does not divulge information about any one individual. There are two common ways to aggregate point data records based on the polygons they fall in: desktop GIS software and GIS web services.

Desktop GIS Software

When using desktop GIS software, the user would add a layer for both the point features (records after they have been geocoded) and the polygon features (boundaries of the area features to use as the units of aggregations). A GIS tool like *Identity* in ArcGIS or the *Join Attributes by Location tool* in *QGIS* will add the ID and/or other attributes from the polygons to the points that fall in them. The point records can then be summed to provide a count per polygon ID. This sequence of steps can be scripted to automatically update on an appropriate cadence.

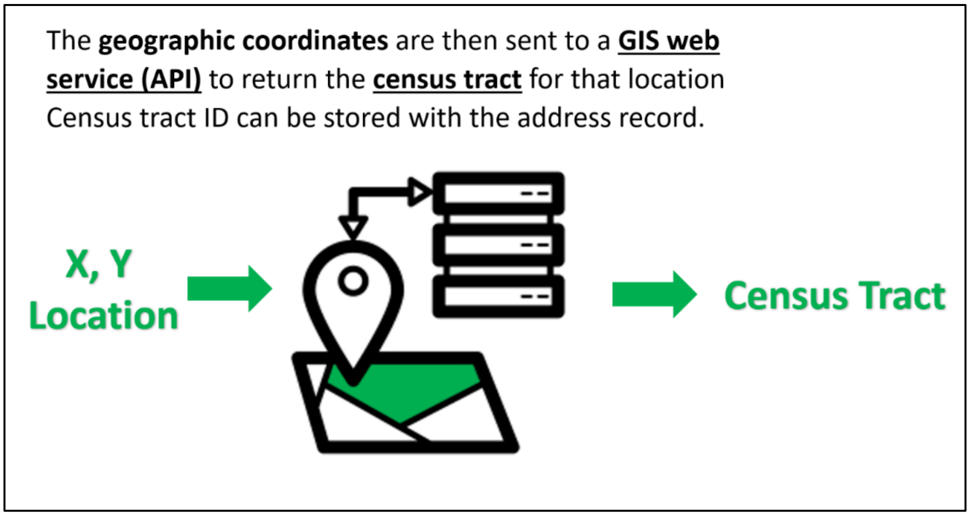

Figure 5: Step 2 is to drop the Longitude Latitude coordinates onto a polygon. This is how you wind up with a census tract.

GIS Web Services

Any polygon base map layer can be published as a GIS map service. The URL for that service can accept a coordinate pair and return the attributes of the polygon that the coordinates land in. This is really what is happening when you click your mouse on an interactive web map to get some information. This function can be done behind the scenes either in batch or as records are entered. For example, using the *GetFeatureInfo* request, a developer can retrieve the attributes of a polygon containing the coordinates passed to an Open Geospatial Consortium web service.

A **table** containing the **total count by census tract** is incremented for that census tract returned

Census Tract	Count	Time and Date Stamp
36081044800	21	3/15/2021 08:12:20 AM
36081045800	45	3/15/2021 08:17:00 AM
36081046200	90	3/15/2021 09:00:00 AM

Census Tract	Vaccination Count	Time and Date Stamp
36081044800	32	3/16/2021 09:18:58 AM
36081045800	53	3/16/2021 10:12:00 AM
36081046200	105	3/16/2021 11:04:00 AM

Figure 6: Step 3 Once the information is dropped onto a polygon, it creates a census tract.

Privacy and HIPPA compliance are concerns when the counts in a geographic area are exceedingly small, so it is common to set a minimum threshold. Maps created from these counts, and statistical analysis using them, can be noted to reflect this. Using the right units of geography as the polygon takes some consideration since the smaller the geographic unit, the more localized the analysis can be.

Census Tract	Vaccination Count	Time and Date Stamp
36081103900	~~1~~ 5	3/17/2021 10:01:58 AM
36081104700	7	3/17/2021 11:11:06 AM
36081027800	~~2~~ 5	3/17/2021 11:24:12 AM

If Count > 0 and <=5, set to 5.

Figure 7: Step 4 To protect privacy when too few people are in the tract, anything between 1-5 can be counted as to avoid reproducing any information about an individual.

Spatial Resolution

The word *resolution* in this case refers to the usefulness of a data set or geographic area to resolve differences at various geographic sizes. Resolution is related to map scale. Small scale maps cover larger geographic areas with less detail than large scale maps. If you were planning a cross country road trip, you would not start with a map of your town. Likewise, the package delivery driver is not looking at a map zoomed out to the whole US to deliver the package to your doorstep. These concepts are intuitive and need to be considered when deciding on the spatial resolution appropriate for pandemic data. Like the trip planning or delivery use case, the resolution decision must be rooted in the task at hand.

Spatial resolution decisions are also influenced by the characteristics of existing data. For example, polygon data sets exist and are readily available for the outlines of nations, states, counties, and municipalities. Further, the US Census Bureau starts with the small units of geography called census blocks. Census blocks are formed by the intersection of streets, railroads, municipal boundaries, and water features. Blocks are then aggregated to form block groups. Block groups are further aggregated to form census tracts. Tracts are optimal areas with a population of about 4,000 people but can vary between 1,200 and 8,000 people.

Zip Codes are often incorrectly thought of as polygon areas, but as defined by the United States Postal Service (USPS), they are not. The USPS does not publish an authoritative set of zip code polygons, but many private companies have created their own approximation of zip code polygons. Zip codes do not cleanly overlay over any other polygon geography and should not be used for data analysis.

Not only should you consider the existence of readily available polygons for data aggregation, but you should also consider the other data available at those same units of geography that might be helpful in future analysis. The Census Bureau published a wealth of demographic data by census tract.

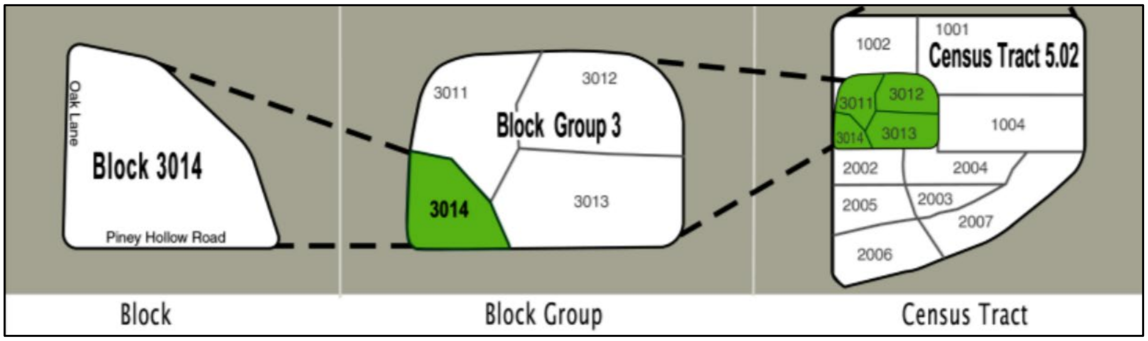

Figure 8: An example of data available within the same units of geography.

Temporal resolution

Values of making data public:
-Transparency/Public Trust
-Removes barriers to sharing
-Unanticipated contributors
-Much simpler technology

Like spatial resolution, temporal resolution is a measure of granularity indicating the appropriateness of a data to resolve differences in elapsed time. Both frequency of update and the lag time are factors in temporal resolution. Consider the time passed between a patient becoming infected and their positive test showing up as part of a daily report. Now consider time passed between an infection and its indication as part of a wastewater surveillance sample.

The time between when something happens needs to be considered. In the case of an infection, the time elapsed from contact to the specimen collection to the date report, to the date the report is added to an information dashboard or analysis. When including any data in reports, it is important to stay consistent. Changing reporting from daily to weekly will create spikes in dashboards.

Data practices

GIS data need to be connected to health system data. A key strategy is to spatially enable and connect to existing systems instead of recreating them in your own GIS system. When designing the connections between healthcare systems and GIS, it is critical to not reinvent the wheel. GIS practitioners need to be aware of existing healthcare data standards including HAVE 1.0, HAVE 2.0, HL7 V2, TEP. GIS efforts should focus on the spatial location of facilities and include unique IDs to join in data from healthcare data systems to spatially enable it. Automatic system to system integrations is critical. Too often, data systems are connected by staff making extracts of databases and sending them as email attachments to be included in other applications and dashboards.

With good planning, data problems that will occur during a pandemic that will impact GIS work can be anticipated. The inequities of pandemic impacts have been well documented. For example, racial and ethnic variations have been highlighted. Unfortunately, race and ethnicity are not high-quality data points reported by laboratories to public health. Jurisdictions have reported missing race and ethnicity data in more than 50% of labs. Prior to the pandemic, jurisdictions should review laws to ensure these data elements are reported and begin active outreach to labs that do not report. During the pandemic, public health can enrich missing case data by looking up cases in other databases (e.g., IIS, hospital EMR). Further, proper data aggregation can provide a direct comparison between a public health phenomenon and high quality racial and ethnic characteristic of a geographic area.

Hospitalization and death record cleanup
In most jurisdictions, hospitalizations or deaths from reportable diseases are not mandated or usually reported to public health. Public health usually only sees that a lab has reported a new case but is unaware if the person is hospitalized or has died from the illness. As a result, hospitalizations and deaths reported are usually delayed and underreported. Manual efforts to look up cases in hospital EMRs and death record searches can improve data quality, but without significant informatics improvements to advance electronic case reporting, these events will remain an issue for the next pandemic.

Federal partners do not report individual vaccination reports to state Immunization information systems (IIS). Only aggregate counts are provided to the CDC. As a result, jurisdictions with large federal workforce populations will have incomplete vaccination data to help communicate community vaccination progress and identify gaps.

While residents and staff of long-term care facilities were rightfully prioritized for vaccination during this pandemic, unfortunately, the vaccination administrators did not universally update the entire IIS record to reflect that the person lived in a long-term care facility (i.e., change their address). As a result, facility vaccination rates could not be calculated using the IIS but had to be gathered manually from each facility.

Unlike other countries where an age-based priority group system was used to rollout the vaccine, the federal government chose to employ multiple pathways for persons to qualify for vaccination (e.g., age, health status, employment). These groups were not able to be captured in the IIS. As a result, many jurisdictions did not ask for proof (e.g., hospital ID) and allowed many persons to incorrectly jump the line.

Credits

We are grateful for the hard work and participation of members of the National Pandemic GIS & Informatics Task Force. The Task Force started as a collaboration between major not-for-profit organizations in the geospatial industry: the National Alliance for Public Safety GIS (NAPSG) Foundation; The Urban and Regional Information Systems Association (URISA); and the National States Geographic Information Council (NSGIC). As the Task Force matured, the National Association of County and City Health Officials (NACCHO) and the Association of State and Territorial Public Health (ASTHO) joined the effort.