CCRA® Training Program

Geocoding & Hazard Retrieval

October 4, 2019

Geocoding and Hazard Retrieval

This CCRA course provides detailed information on the geocoding and hazard retrieval processes in catastrophe modeling. The course will include the following units:

Unit 1: Geocoding and Impact on Loss Results

This unit will focus on the basics necessary to better understand the geocoding process and may cross reference materials from other courses.

Learning Objectives:

- Demonstrate an understanding of the analytical context of geocoding.
- Describe some of the properties of geocoding accuracy.
- Understand the consequences of data quality for geocoding.
- Enumerate some of the downstream effects of geocoding on other model components.
- Understand the drivers of change in loss due to changes in geocoding.
- Explain variability of loss between and among geocoding levels, based on the resolution of specific perils.

Specifically, we will cover the definition of geocoding, the relevance of geocoding for modeling, geocoding accuracy and uncertainty, data quality challenges, downstream effects of geocoding, how the geocoder finds a match, what the geocoder does with match results, and geocoding's impact on loss results.

What is Geocoding and Why Does it Matter?

Geocoding is the process of translating address data (street address) to global coordinates (latitude and longitude). Peril models cannot recognize addresses, such as 123 Main Street; they require X-Y coordinates (latitude and longitude) in order to perform the geographic calculations necessary to ultimately generate loss. Address information is validated using pre-compiled databases of known deliverable street addresses, valid postal codes, etc. The geocoding process enables the software to link building address information with tabular data in the models, and is one of the first steps in catastrophe modeling. As we will see, the geocoding process and the positional accuracy of individual locations are linked to the input data.

In addition to translating address data, geocoding validates the address elements, such as postal code or city name. During the geocoding process, these address elements are compared to an internal reference list and standardized to facilitate the most accurate information for the modeling process. Geocoding also allows complex spatial operations to be performed, such as distance calculations or point-in-polygon queries, and "back-filling" of other data elements that the modeling software uses for look-ups.

The accuracy of geocoding and the understanding of the geocoding results are important. Many business decisions are made based on geocoding results and, ultimately, loss estimates. But decisions made based on geocoding are only as sound as the data that support them. The quality of data that are input into RMS catastrophe models influences what hazard information is retrieved, the accuracy of the analysis by a catastrophe model, and the accumulation results. The importance of data resolution and data quality should not be underestimated. As the geocoding resolution increases, the number of assumptions made during the modeling process decreases.

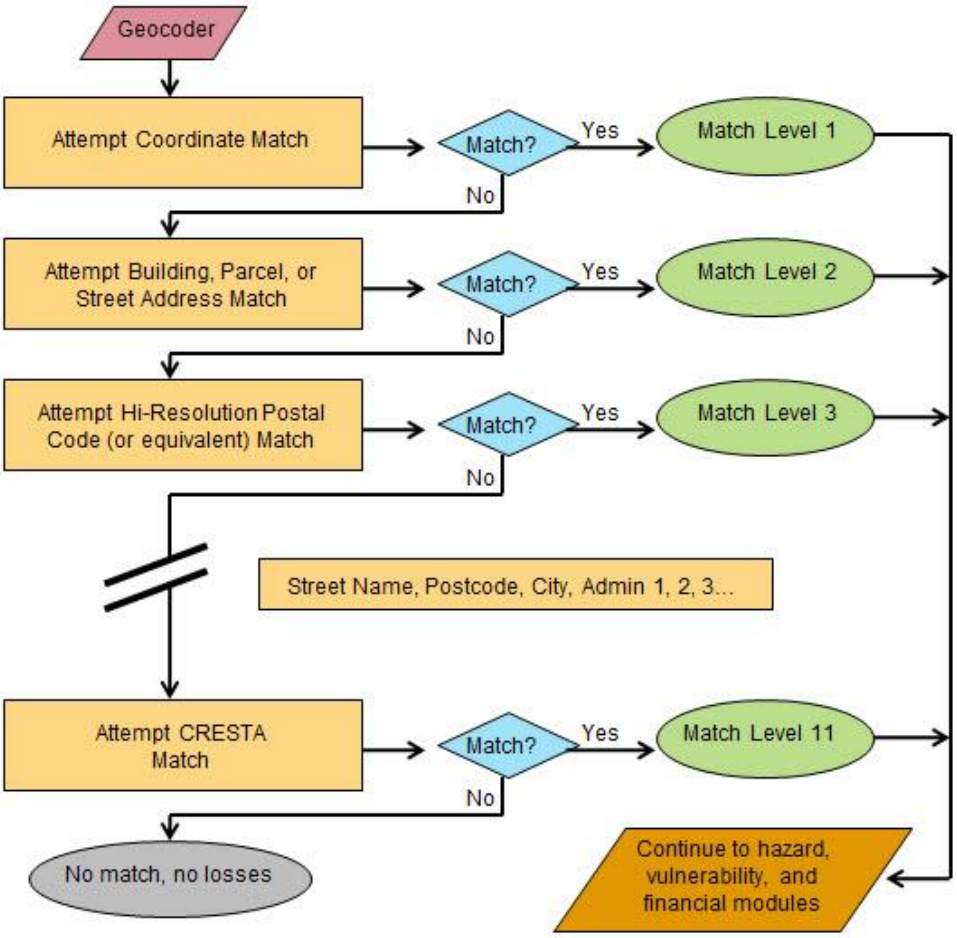
The Geocoding Process

How does the geocoding process work? Geocoding outcomes can vary from pin-pointed coordinates to large regional units (e.g., CRESTA zones). Geographic information in different countries varies in terms of quality, format, and resolution. Therefore, the geocoder uses a series of standard protocols globally to match geocode information for the best result possible.

After the address is entered by the user, geocoding software attempts a match at the highest resolution possible – the coordinate level. If this succeeds, then other geographic data are retrieved and stored (e.g., city code, county code). If this fails, the software then attempts a match at the next lower resolution, which is a building, parcel, or street address level. If this succeeds, then other data are retrieved and stored. If this fails, then the software attempts a match at the next lower resolution until the lowest resolution is reached.

If the geocoder is not able to match at any level, a result of none or 0 will be returned, and the location is excluded from subsequent analyses.

FIGURE 1: Recursive Address Matching Diagram

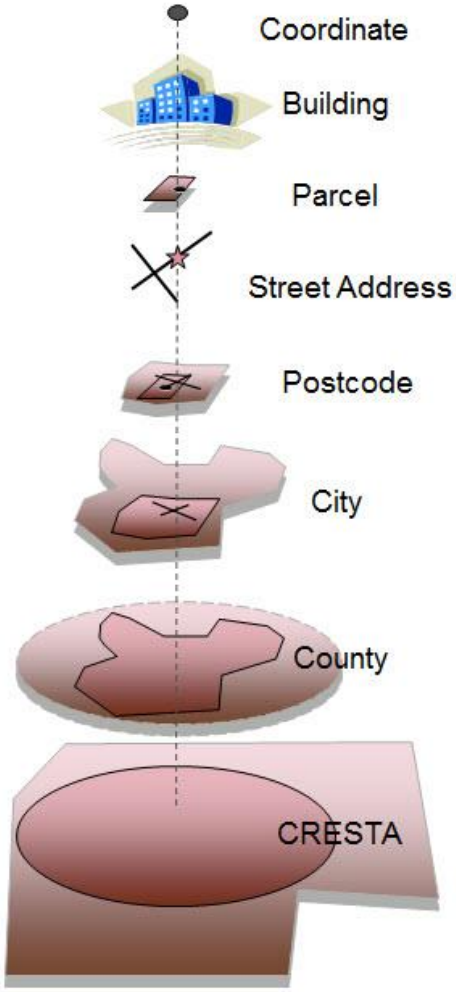


E-LEARNING INTERACTION 1: Recursive Address Matching

To view an interactive example of the recursive address matching process, go to the Geocoding and Hazard Retrieval course in the CCRA portal of Owl and select Interaction 1: Recursive Address Matching.

Given the wide range of address information quality, there is a range of potential geocoding resolutions. The geocoder may be able to recognize the exact location, or it may only be able to identify the postal code. In the latter case, the geocoder may have to assign a centroid of the postcode to the location. The match level achieved when geocoding depends on the quality of the address information and the available geographic data in the geocoding engine.

FIGURE 2: Geocoding Match Levels U.S.



In the United States, there are some geocoding match levels not generally available elsewhere in the world, which include the building and parcel levels. In RMS products, the building level match uses RMS High Resolution Buildings (HRB) data, which include points representing the centroids of four-wall structures that may have multiple addresses. Parcel data, which will be discussed in unit 2, are compiled from a variety of sources and are linked to the extent of the property. While parcel centroids generally provide good positional accuracy, especially for small residential properties, there is no guarantee they will fall on the actual location of the buildings within the property.

Outside of the U.S., geocoding match levels for a high resolution postcode or block, such as a U.K. or Canadian 6-digit Local Delivery Unit (LDU) point, usually match to within a block or better of the true location. Another RMS match level used outside the U.S. is the Admin3, commonly used to identify a district or similar resolution. In some countries, the Admin3 is considered higher resolution than a city, such as the Mexican Colonia. In other countries, Admin3 may be a lower resolution than a city, as in the German Gemeinde.

E-LEARNING INTERACTION 2: Geocoding Match Levels

To review an interactive example of the geocoding match levels, go to the Geocoding and Hazard Retrieval course in the CCRA portal of Owl and select Interaction 2: Geocoding Match Levels.

Geocoding in Detailed Versus Aggregate Analyses

The Quality

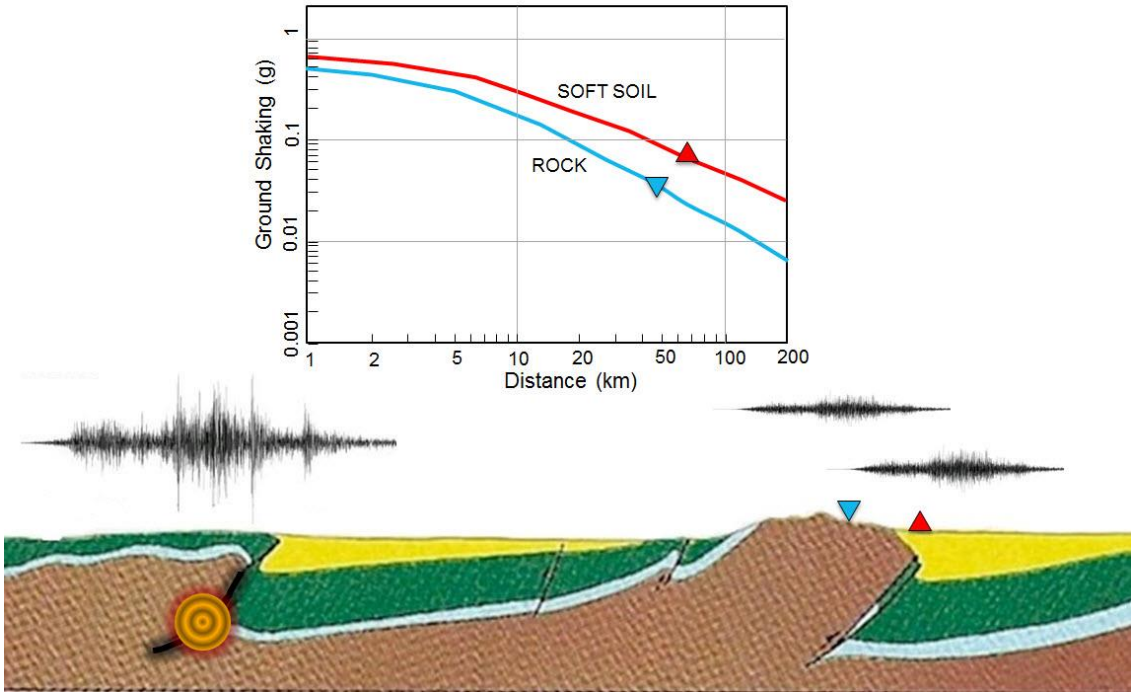
One important component of a catastrophe model lies in the granularity of the loss calculations: are they performed for each individual location and later combined, or on an aggregated group of locations? This impacts the structure of the model and its underlying assumptions. In this section we discuss the RMS differentiation between these two, but the principles are relevant in understanding other models. Aggregate and detailed modeling will also be discussed in the Financial Modeling course.

For every country peril model, RMS offers a detailed analysis solution via RiskLink®-DLM (Detailed Loss Model) and an aggregate analysis solution via RiskLink-ALM (Aggregate Loss Model). The level of address resolution obtained and the model's resolution will impact whether a DLM or an ALM analysis is the most appropriate.

The DLM uses the geocoder and relies on precise geographic information about each location. It is based on factors that affect a particular site. Because site hazards can vary over very short distances, precise location information results in more accurate hazard information. A detailed analysis accounts for the specific conditions at an individual site or location.

For example, Figure 3 shows that source energy from earthquakes affects large areas. This energy degrades, or attenuates, with distance. The degree of ground shaking may be altered by local conditions at the site of a particular property. Ground shaking can be amplified or dampened depending upon variables like soil type. The locational accuracy for a site affects the model's representation of both its distance from the earthquake and what local conditions are assumed.

FIGURE 3: Example of Detailed Earthquake Modeling

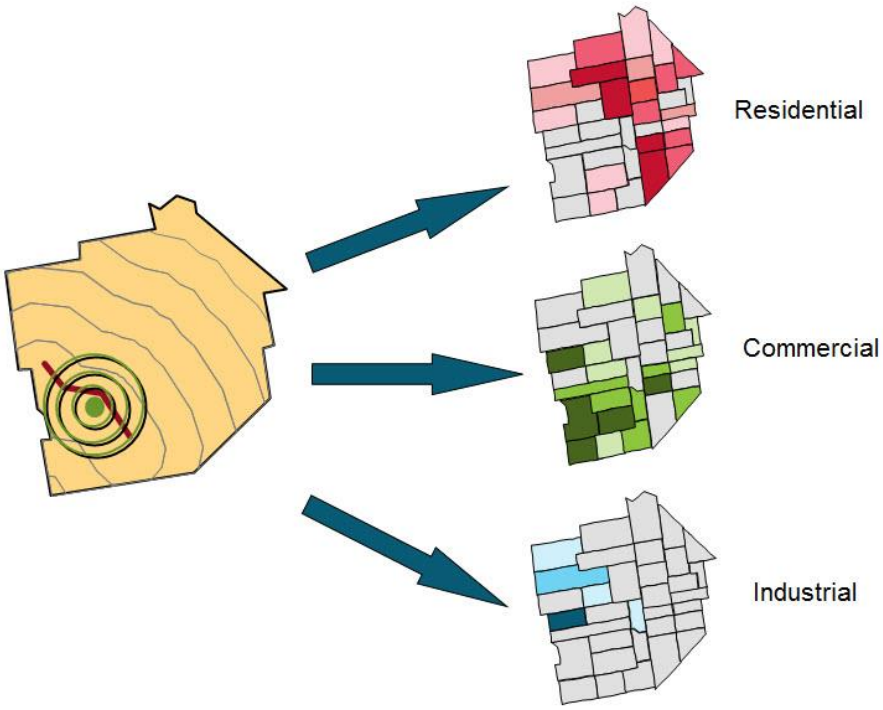


Aggregate modeling is used when we cannot identify a property's precise location, or if an analyst chooses to group data in order to obtain results in a shortened period of time.

The ALM is much less reliant on the RMS geocoding engines than the DLM, as the ALM analysis is conducted on limits or sums insured within larger geographic units, such as CRESTA zones. The ALM is based on precompiled analyses in which results for an artificial portfolio at higher resolution have been weighted by exposure. The exposure weights and policy assumptions vary by line of business.

It is important to understand that using these assumed distributions of value can lead to differing conclusions about risk. As the quality of address information degrades, the model relies increasingly on pre-compiled distributions, which represent industry averages. These industry-based assumptions may not match your mix of business or policy conditions, and as a result, your loss result may not be as reflective of reality as it could be. The best way to obtain an analysis that is reflective of your portfolio's unique characteristics is to provide the model with detailed information on each property in the portfolio. Aggregate analysis still allows for successful analysis, but if the exposure is not similar to the assumed industry distributions, the resulting analysis results may not be representative of the actual risk.

FIGURE 4: Assumed Distributions in Aggregate Modeling



In summary, there are some important differences between ALM and DLM. First, DLM is site specific, whereas ALM uses collections of exposure by geography, such as an entire CRESTA zone. Second, in DLM, losses are estimated using detailed geographic information that may cause losses to be highly variable over short distances. In contrast, ALM uses pre-compiled data that are averaged over larger areas, depending on the geography that is being used (e.g., CRESTA). Lastly, DLM loss results can be saved down to the individual location level, whereas loss results from ALM are representative for the area of the analysis only, not for an individual location.

Here are some general rules of thumb to keep in mind when analyzing results from either the DLM or the ALM model:

- The more accurate and more detailed the information input into the model, the better the spatial accuracy of the results.
- As geocoding resolution improves, modeling uncertainty decreases.
- Detailed analysis happens when we have enough detail on each location to calculate loss on a site by site basis.
- When there is not enough information about a location we can rely on an aggregate analysis, which allows assumptions to be applied regarding the likely distribution of properties in an area.

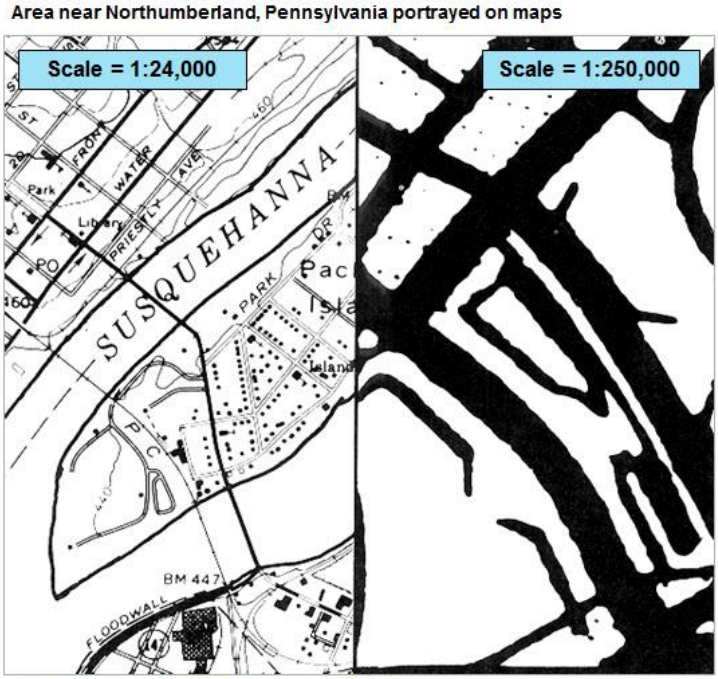
- Geocoding at the postcode or street-name resolution is often sufficient for DLM loss analyses, but high-resolution geocoding (at the street address or better) produces more realistic, modeled results.

Map Scale and Spatial Accuracy

The visual aspects and ease of zooming into digital maps can provide a false sense of the accuracy of the underlying spatial data. The uncertainty embedded in paper-based technology provides a means to understand both the strengths and limitations of geocoding. Some of the data behind geocoding and hazard retrieval still originate with paper maps and, as a result, contain both human and technological limits. The single most significant limitation of paper-based source information is the thickness that a line may take in order to represent a feature or a boundary.

The following figure shows maps at two different scales. The map on the left is at a scale of 1:24,000 and shows a small area in much detail. You can see the text Susquehanna River, a railroad, street names, and even buildings. The map on the right is originally a 1:250,000 scale map, covering a large area but consequently in much less detail. This map is magnified to match the same area on the Susquehanna River as the 1:24,000 scale map at the left. The 1:250,000 map is very crude in the level of detail in comparison. The lines representing roads and boundaries were roughly the same thickness on the original paper map, but the enlarged lines show only a small fraction of the information that easily fits into the larger scale 1:24,000 map. The lines are also coarser, so when digitizing to create a digital map, there is a greater margin for error when determining where the street line is actually located.

FIGURE 5: Map Scale Challenges – Portraying Maps at Different Scales



Source: M. Monmonier, *How to Lie with Maps*

The U.S. government recognized that map scale has a direct impact on spatial accuracy and so established federal standards in the 1950s to certify and maintain information quality in published maps. All published maps must be accurate within a specified range. For a 1:250,000 scale map, a line can be no more than 694 feet away from its true location and a point feature no more than 208 feet away. Much of RMS' hazard data layers are derived from paper maps, either digitized by RMS or the issuing agency. Paper maps are primarily issued by government agencies. So even with these standards, there can still be an error of almost 700 feet on a 1:250,000 scale map.

These problems are compounded by the variability inherent in hazard data (e.g., soil type and flood zones). It is still technologically impossible for a single line to capture a zone of transition between soft or hard soil, for example. A single line also cannot accurately capture the boundary between a 100-year and 500-year flood zone. Most geographic information systems (GIS) represent such data layers with vector-based polygons, in which it is possible to keep zooming in on a boundary far beyond the scale at which that boundary remains meaningful. Understanding scale is critical to appreciating one of the most significant sources of error that can be introduced into accumulation and peril modeling.

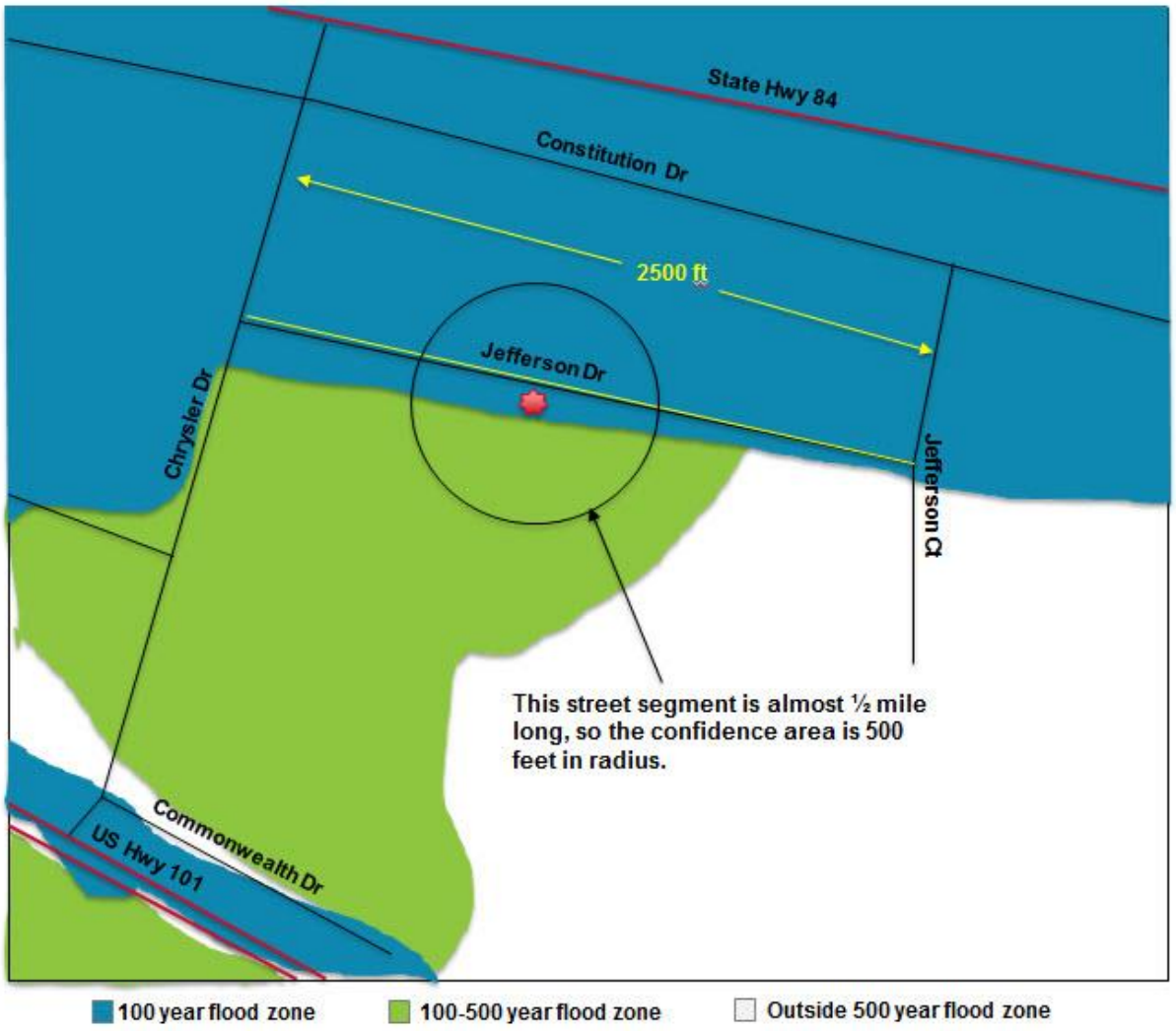
Visualizing Spatial Uncertainty

Spatial accuracy in modeling has multiple influences, including the resolution of geocoding base maps, the precision of geocoding methodology, and the resolution of hazard information. Uncertainty arises from the creation of both paper-based maps and digital maps. During the creation of digital maps, conversion errors can occur from scanning paper maps and human errors can occur during the digitization process. Potential errors can also occur in the location and ranges of addresses introduced by the vendors who create the databases.

RMS occasionally uses a buffer, or what we call confidence areas, to account for accumulated uncertainty from all these sources, and to be able to express that uncertainty to end users.

Looking at the graphic in Figure 6, the circle represents the confidence area. The red dot represents the best available estimate of the location. It appears to be within a 100-year flood zone, but because of accumulated uncertainty, the location could actually lie anywhere within the confidence area, which could place it in a 500-year flood zone. This demonstrates why it is important to consider the confidence area information provided in the lookup results.

FIGURE 6: Visualizing Spatial Uncertainty Using Confidence Areas



RMS determines the size of the confidence area by the length of the matching street segment. The longer the street segment, the greater the uncertainty as to the actual position of a location, and therefore, the larger the confidence area. If a street level match or better is achieved, RMS calculates the confidence area equal to 20% of the length of the matching street segment and stores this value in the GeoBuffer field.

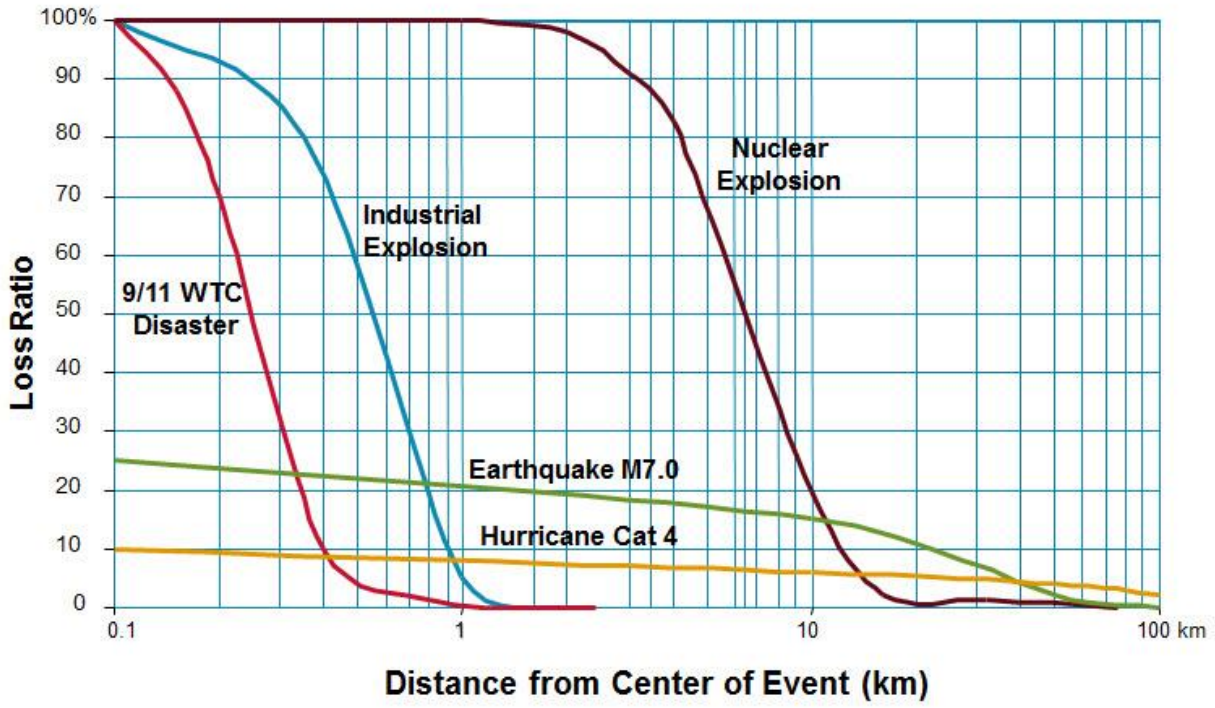
The confidence area is not used in modeling uncertainty calculations or in the loss results. However, it is useful to understand the spatial uncertainty associated with a high resolution geocoded location, especially when evaluating a location with respect to flood zones as indicated in Figure 6.

Spatial Accuracy in Peril Modeling

Loss curves can vary significantly according to characteristics of the peril and affected exposure, but as a rule of thumb, loss ratios degrade over space outward from the center of any disaster. An awareness of the degree of variation is useful when considering the relative importance of geocoding accuracy.

The following figure is a highly simplified comparison of the decay in mean damage with distance for several perils. The steeply curved red line illustrates the decrease in loss from the September 11, 2001 terrorist attack at the World Trade Center. Losses were total at the site of the attack, but rapidly decreased with greater distance from ground zero.

FIGURE 7: Radius of Loss for Engineered Commercial Property from Center of Disaster



With natural catastrophic perils such as earthquakes and hurricanes, losses are generally lower overall at the center of the disaster when compared to man-made catastrophes. However, losses from natural catastrophes tend to drop more gradually with distance than what is typically seen with man-made disasters, such as terrorist attacks.

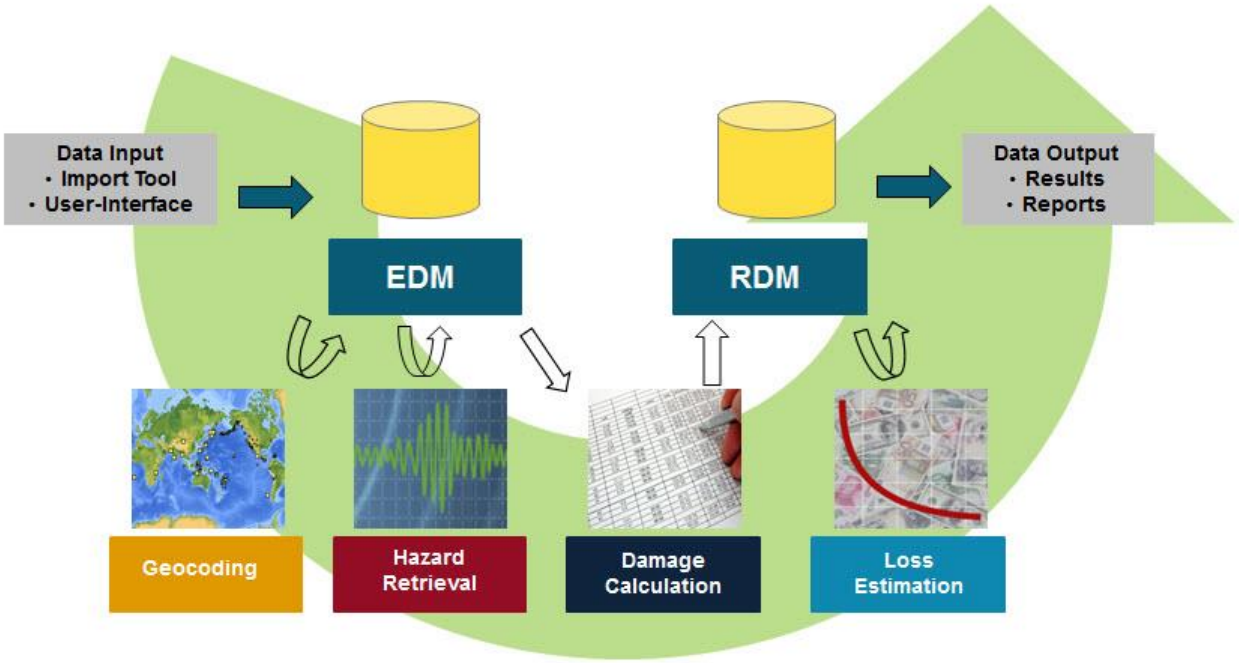
Spatial accuracy is therefore more critical for catastrophes with steeper loss curves because there is high variability in damage within a smaller geographic area.

Downstream Effects of Geocoding

Geocoding is the first modeling step within RMS software and it affects all subsequent components, including hazard retrieval, damage calculation, and loss estimation. The address

match level determined during geocoding defines the level of resolution of the hazard lookups and the associated analyses. The following diagram illustrates how data flows within RMS software applications.

FIGURE 8: RMS Software Components



In RiskLink and RiskBrowser, location and exposure data are entered by the user and immediately sent to the Exposure Data Module (EDM) database. This information is then sent to the geocoder, where the location data are enhanced with the geocoding results and sent back to the EDM. These data are then sent through a hazard retrieval process, where the geocoding information is used to perform geographic lookups on RMS hazard data layers for relevant selected perils. All of these data are passed back to and stored in the EDM, where they impact modeling components such as damage calculation and loss estimation or may be used on their own for underwriting decisions.

Users have the most influence right at the beginning of the process – with the quality of address data that is input. There are several downstream effects of geocoding, including **hazard retrieval, inventory selection,** and **vulnerability curves.**

First, **hazard retrieval** is the function most directly impacted by geocoding. Hazard retrieval includes a variety of actions on a location by a given model, including the selection of relevant events, calculation of the metric used for damage calculation, and assignment of site-specific characteristics that can affect the base hazard. The magnitude of the differences due to potential geocoding results vary by peril. Earthquake results, for example, can show sharp gradients in hazard with distance from a fault and are heavily influenced by the site level

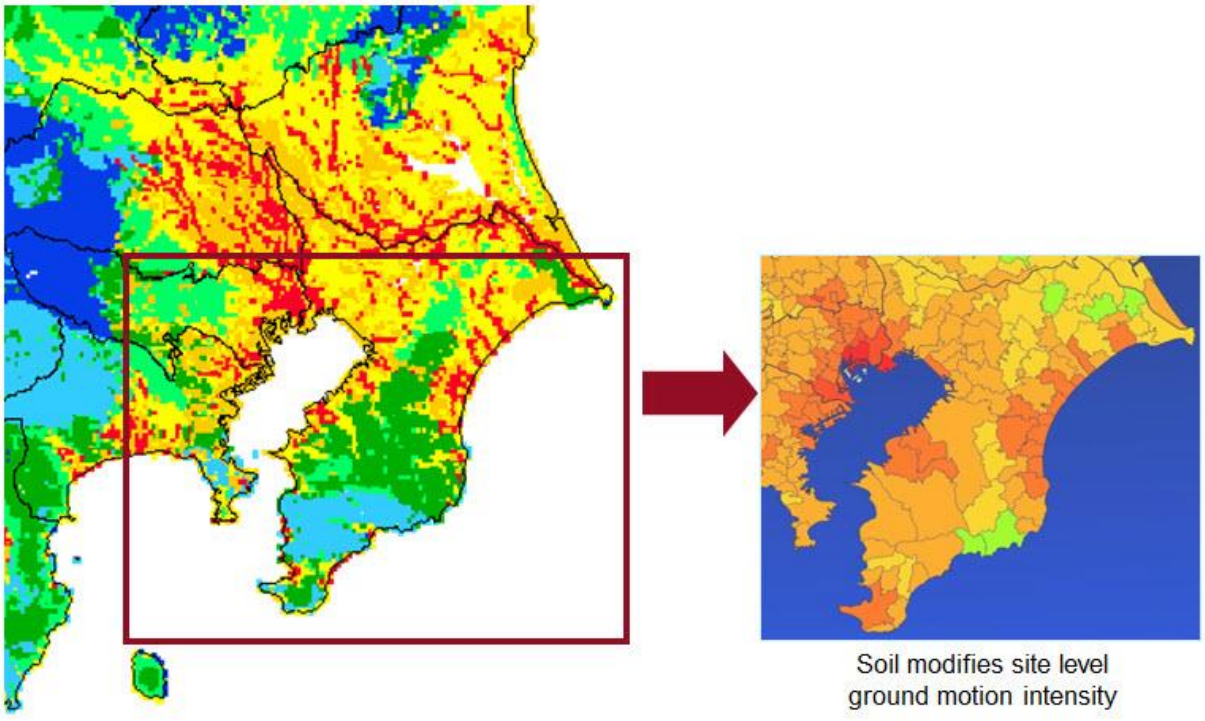
features (e.g., soil). Windstorm, severe convective storm, and flood results are more event-based, meaning geocoding is more relevant in how events are selected.

The connection between event selection and geocoding resolution may not be immediately apparent, but it is important. The model will find events that affect your location using a Variable Resolution Grid (VRG) cell. The VRG is discussed in detail later in this course. For this section, however, it is important to know that it is a global grid system developed by RMS where the grid cells are of variable size depending on the risk and exposure level. Using hurricane as an example, in high risk and high exposure areas (e.g., Miami, Florida), the grid cells can be as small as 50x50 m. In low risk areas with little variation in hazard, grid cells can be as large as 100x100 km. RMS creates and stores hazard data at the VRG level, and identifies which events impact each cell. Lower resolutions are modeled with either a coarser VRG cell or a weighted average over the larger area.

As a result, when high resolution geocoding is possible, you get the most accurate selection of events that will impact your location(s). When geocoding at the postcode or admin2 (e.g., county) level, the model will choose a much larger set of possible events as these larger areas will impact multiple VRG cells in most cases. The end result is an increased uncertainty in the loss estimates.

An example of how geocoding impacts hazard retrieval for the earthquake peril is shown in Figure 9. Lower geocoding resolution (i.e., coarser resolution) leads to the use of aggregate hazard data, which increases assumptions and uncertainty. This slide shows two maps of Tokyo, Japan. The map on the left shows soils for the Tokyo area at a high resolution (1km^2 grid). This map has much more soil detail than the map on the right. The map on the right shows soil at the city/ward level, with values representing the average conditions over the area. Higher geocoding resolution allows the model to access more site-specific hazard data, which will result in more realistic loss results.

FIGURE 9: Impact of Geocoding on Earthquake Hazard Retrieval



One rule to remember is that hazard information cannot be retrieved at a higher resolution (i.e., finer resolution) than the geocode match level. For example, if you get a city level geocoding match, you will not see a corresponding postal code level hazard lookup. The implication of this is that the geocoding resolution impacts the resolution of the hazard data retrieved. The geocoding resolution, therefore, will also impact the loss estimates for the entire portfolio.

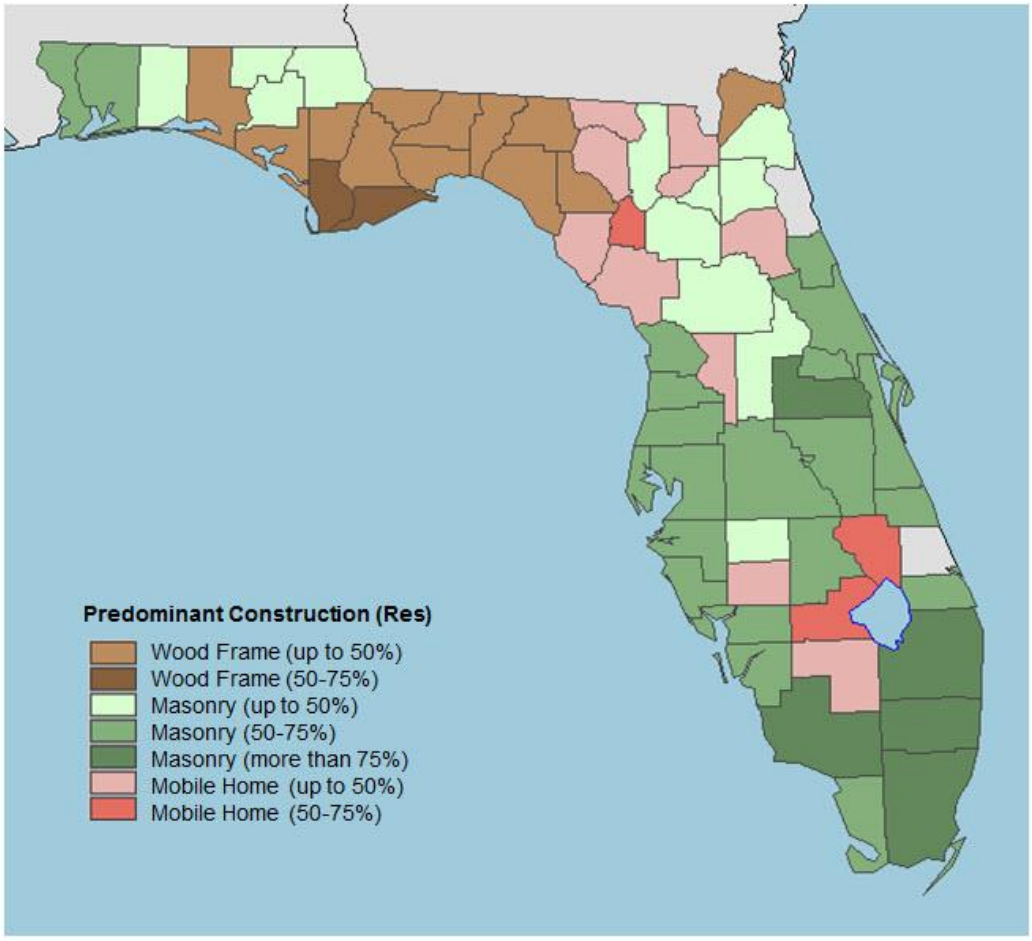
While hazard is typically the model component most significantly impacted by geocoding results, vulnerability assumptions can also be affected. Foremost among these is the **inventory selection**, which becomes relevant when detailed construction data is not entered by the user. If construction type and occupancy are unknown, the model will choose a pre-defined building mix based on the geocoded data.

Unknown data fields, such as construction type or year of construction, force the software to make assumptions about the average property in the area. RMS provides inventory mixes that vary within a region, as shown in the map in Figure 10. Inventory resolutions vary by model and region, including admin2, admin1, VRG, or postcode depending on the data available and degree of variation.

The following map shows residential inventories at the county level in Florida. It also shows the predominant type of single family construction by county. Note that some of these counties have 50% or more mobile homes. An unknown construction type in one of these counties would cause a property to appear more damageable than it really is if, in fact, it is a single family home

made of brick veneer or masonry. This map shows an example of what happens when you do not have detailed site information and must rely on pre-compiled information.

FIGURE 10: Effect of Location on Inventory Selection for Florida Residential Portfolio



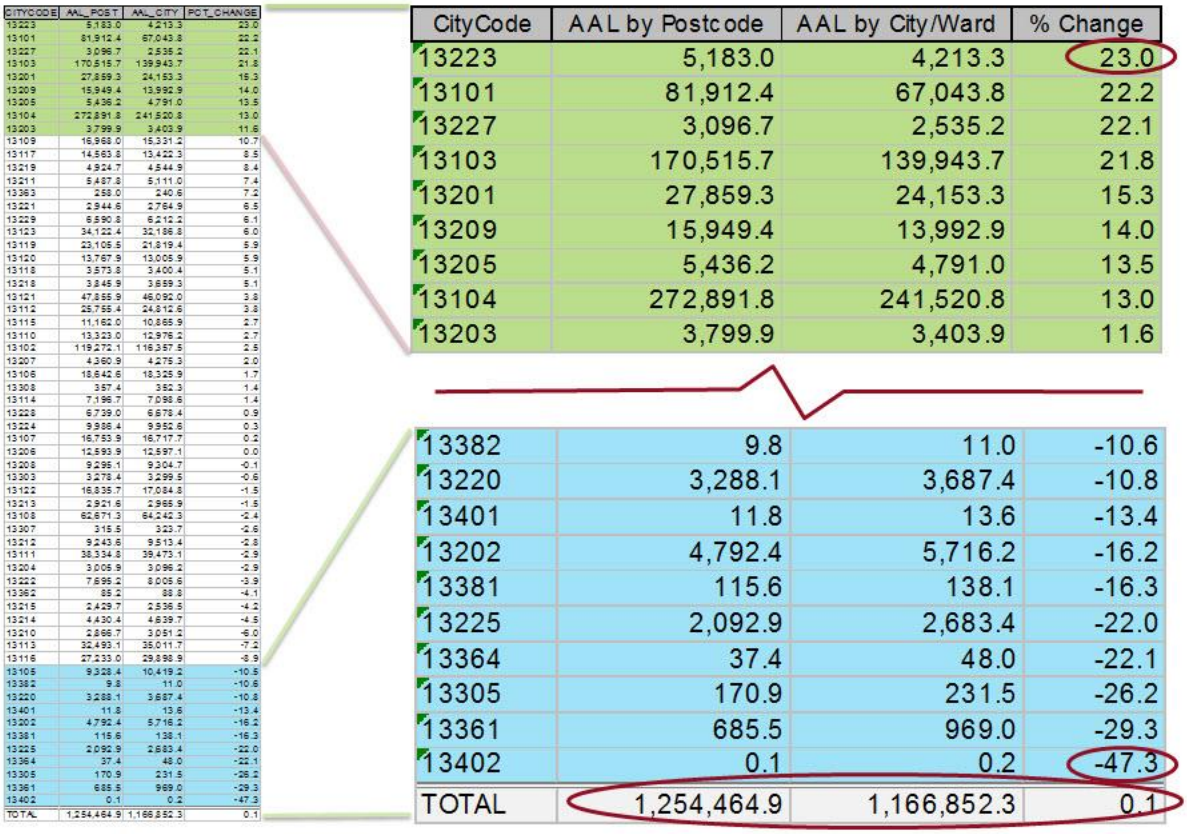
Lastly, **damage curves** can also be affected by the geocode results for a location because **vulnerability** regions are defined geographically. In the U.S., a postal code may cross county boundaries, but it may be assigned to only one county. Depending on where the property is actually located, geocoding results may assign that property to the “other” county, and if the two counties lie in different vulnerability regions, then the geocoding match level directly affects the damage curve selection. Another example occurs in Florida, where a different set of damage curves are used for locations within half a mile of the coast.

Case Study: Geocoding Impact on Japan Earthquake Analysis

Using a case study of a Japan Earthquake analysis, we will look at how the differing geocoding results affect the average annual loss results for a portfolio.

The right side of the figure below shows two sections from the table shown on the left. The tables on the right show those city/wards with the largest difference in average annual loss (AAL) when locations are geocoded at the 7-digit postcode vs. at the city/ward level.

FIGURE 11: Japan Earthquake Analysis Average Annual Loss Results



CityCode	AAL by Postcode	AAL by City/Ward	% Change
13223	5,183.0	4,213.3	23.0
13101	81,912.4	67,043.8	22.2
13227	3,096.7	2,535.2	22.1
13103	170,515.7	139,943.7	21.8
13201	27,859.3	24,153.3	15.3
13209	15,949.4	13,992.9	14.0
13205	5,436.2	4,791.0	13.5
13104	272,891.8	241,520.8	13.0
13203	3,799.9	3,403.9	11.6
13382	9.8	11.0	-10.6
13220	3,288.1	3,687.4	-10.8
13401	11.8	13.6	-13.4
13202	4,792.4	5,716.2	-16.2
13381	115.6	138.1	-16.3
13225	2,092.9	2,683.4	-22.0
13364	37.4	48.0	-22.1
13305	170.9	231.5	-26.2
13361	685.5	969.0	-29.3
13402	0.1	0.2	-47.3
TOTAL	1,254,464.9	1,166,852.3	0.1

Notice that differences in geocoding resolution may produce small overall differences in loss among multiple locations. However, differences in geocoding can produce significant differences at the site level. The portfolio only showed a 0.1% change in overall loss vs. the first location with a 23% difference in loss between postcode and city/ward level geocodes.

Each of these examples highlights how the geocoding resolution for a location can influence the modeling assumptions. That being said, here are a few important points to remember with respect to geocoding resolution and modeled loss:

- Lower uncertainty (from higher resolution) does not necessarily produce lower losses.
- Higher resolution tends to produce more variation in losses as risk is differentiated.
- Average of higher resolution losses tends to converge on lower resolution results, if the portfolio is similar to the industry.

Unit 1: Closing Key Concepts

Peril models cannot recognize addresses as we do; they can only use latitude and longitude coordinates in order to perform the geographic calculations necessary to generate loss. Geocoding is the process of converting and validating local coordinates (street address, postcode, etc.) into global coordinates (latitude, longitude).

Geocoding is important because it can play a major role in business decisions. Accuracy of position of the location, which means knowing where a location is on the earth as represented by a latitude and longitude coordinate pair, plays a key role in geocoding accuracy.

It is important to remember that hazard information cannot be retrieved at a higher resolution than the geocoding level of the location. The geocoding resolution impacts the hazard data resolution.

In general, location-level losses are lower overall at the center of natural catastrophic perils such as earthquakes and hurricanes when compared to man-made catastrophes; however, the loss ratio drops more gradually with distance than what is typical of man-made disasters. Spatial accuracy is more important for catastrophes with a steeper loss curve because there is high damage variability within smaller geographic areas.

Unit 2: High-Resolution Geocoding

This unit will focus on the technology that underlies the geocoding process, specifically software functionality, address standardization and matching, street-level interpolation, geocoding outside the United States, and filling in the geocoding framework.

Learning Objectives:

- Describe the principal components of geocoding technology worldwide.
- Demonstrate an understanding of geocoding data sources.

The most advanced products were developed first in the U.S. The U.S. government makes subsidized address data available and affordable; however, similar technology is now available for a wide range of countries, which we will explore in this unit.

Geocoding resolution varies around the world because the availability of high resolution data varies. The RMS geocoding software incorporates a flexible framework to accommodate all levels of geocoding for all countries, and will continue to evolve as more data become available.

Geocoding Implementation for the U.S.

In the U.S., the geocoding software and data market evolved earlier than for other countries in the world, due largely to the large amount of address data that was made publicly available.

RMS' Geocoding software has the ability to correct, clean, standardize, and spatially enhance address information for use in RMS products. RMS' geocoding performs a process called "conflation" with the U.S. Postal Service (USPS) data and other commercially available street address databases. Conflation is the process of combining the data from two disparate data sources into a single unified data set. This process creates a larger data set that results in more matches and fewer false positives. In the geocoding context, a false positive is when an address receives a geocoding match, but it is in the wrong place.

In the U.S., address standardization is the process of verifying that each address element meets USPS guidelines for a deliverable address. During standardization, the geocoder compares each address to the USPS database and corrects misspellings, dropped address elements and abbreviations, and provides correct city, state and ZIP Code information. The process of cleaning up the addresses helps to improve geocoding efficiency and hit rates.

RMS uses several sources to perform geocoding in the U.S., and offers the following geocoding databases:

- **RMS High Resolution Building (HRB) Data** is an extremely detailed database for over 400 thousand unique building in the U.S. with attributes verified through site surveys and/or orthographic imagery. This is an optional high-resolution level of geocoding in RMS products. It allows for geocoding to the level of a specific building or four-walled structure, even if it has multiple associated street addresses.

- **Parcels Data Set** is part of the RMS premium geocoding address data. The parcel data provide accurate parcel boundary information for the U.S. The parcel coordinates are based on parcel polygons, which provide better spatial accuracy than street interpolation, which will be discussed later in this course.
- **TomTom Street Layer Premium Geocoding Data** is another of RMS' premium geocoding databases. Street address ranges for nearly every street in the U.S. are included in the database, and nearly every street segment is positionally accurate.

When licensed together, these databases complement each other by working in tandem, and result in more geocoding matches and fewer false positives.

Geocoding Worldwide (Non-U.S.)

In the past several years, geocoding capabilities around the world have begun to approach those in the U.S. The "pull" factors for such a change include new GPS technologies, mobile device tracking, and web search capabilities that have increased data availability. There is also the "push" of internal and regulatory requirements for better risk management practices, including accumulation management on detailed location information, more international models with higher resolution capabilities, and, especially in Europe, consistency in exposure data across international borders. The recognition of the importance of data quality continues to drive the need for geocoding capabilities worldwide.

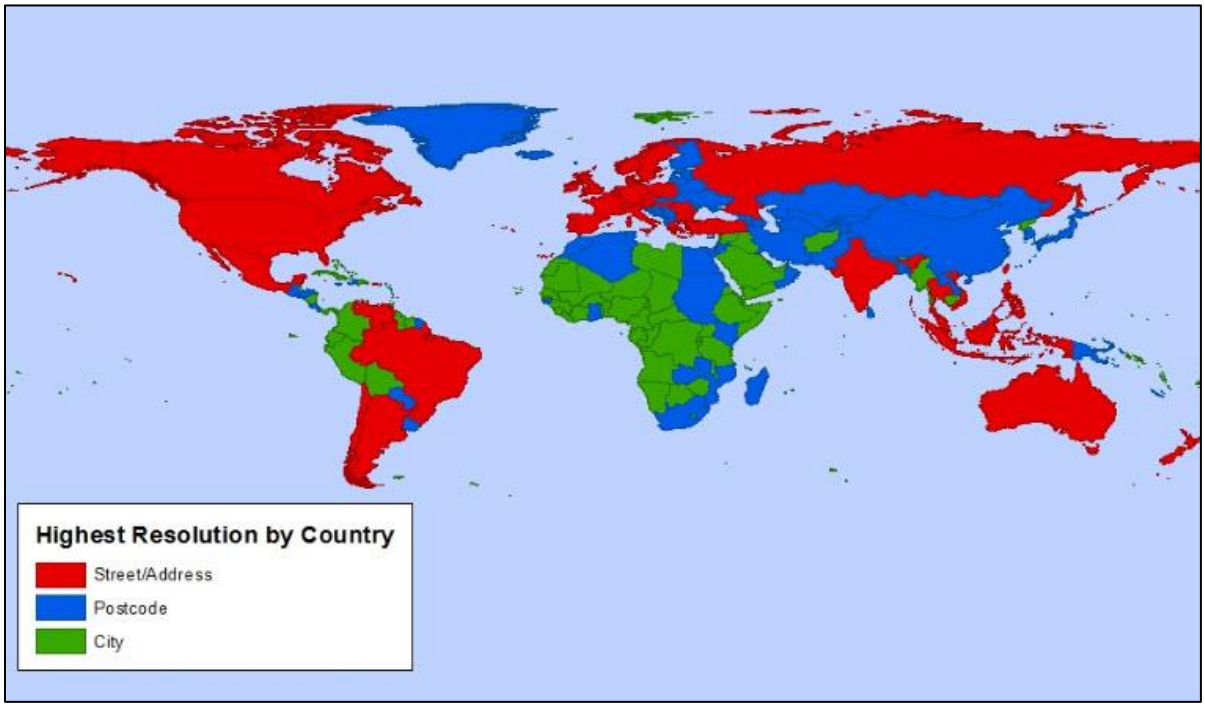
Previously, geocoding technology and data were not available in many countries outside of the U.S., and where they were available, the datasets were prohibitively expensive.

Now the technology exists and street address databases are available for 55 countries. The Global Location Module (GLM) is RMS' geocoding engine, which searches for a geocoding match between the address information input into the model and a list of valid postcodes, cities, and administrative units in the geocoding database. Built from the beginning as an international geocoder, the GLM provides consistency across borders. The GLM works behind the scenes, meaning that it will not be obvious that there is another geocoder working within RMS products, but you will see new and, usually, improved geocoding levels within the user interfaces of these applications.

The GLM has the ability to resolve multiple geographic spellings, misspellings, and common address abbreviations. In addition, it supports special characters and multiple languages. Overall, the GLM supports geocoding for 152 non-modeled countries and 101 modeled countries or territories.

In the following map, the countries in red indicate countries where RMS currently has street level geocoding available. Basic geocoding is available for every country.

FIGURE 12: Representative RMS Geocoding Capabilities



Geocoding resolutions vary by country and RMS will continue to add the best resolutions as they become available and affordable. Catastrophe models run at different levels of resolution and may not take advantage of all available geocoding data.

Still, not every country with high resolution geocoding has the required databases with 100% coverage. The street level data do not necessarily provide a uniform level of resolution for the entire country. Furthermore, not all streets provide address ranges that allow for interpolation. All countries will see increased coverage over time. More detailed data will become available for the countries with high resolution geocoding capabilities, and new countries will be added that will have growing coverage. Data quality and geocoding precision are ever-changing – it will likely never be possible to have 100% coverage for all locations.

Geocoding Software Functionality in the U.S.

Now we will look at how the geocoding process actually works, first in the U.S. There are five steps in the geocoding process:

1. Address normalization
2. Address matching
3. Coordinate retrieval
4. Ancillary data retrieval
5. Join with catastrophe model data

An example of how the geocoding process works in the U.S is shown in Figure 13. Once the address enters the geocoder, it is parsed into its components. Then the geocoding attempts to

find the closet match in the source data. However, the geocoder does not assume that all address elements are entered correctly.

FIGURE 13: Geocoding Functionality in the U.S.

First, note that the ZIP Code was not provided. This was corrected to 94560 through the address normalization process. The word “Boulevard” was updated to “Blvd”, also through the normalization process.

During the address matching, the address on Gateway Boulevard in Newark, California was found, and the latitude and longitude coordinate pair listed here was returned as part of the coordinate retrieval process.

Lastly, ancillary data were appended to the address. In this case, the county and the CRESTA zone were identified and added to the location information.

Once this process is complete, we can join this geocoding data with all of the data required for modeling.

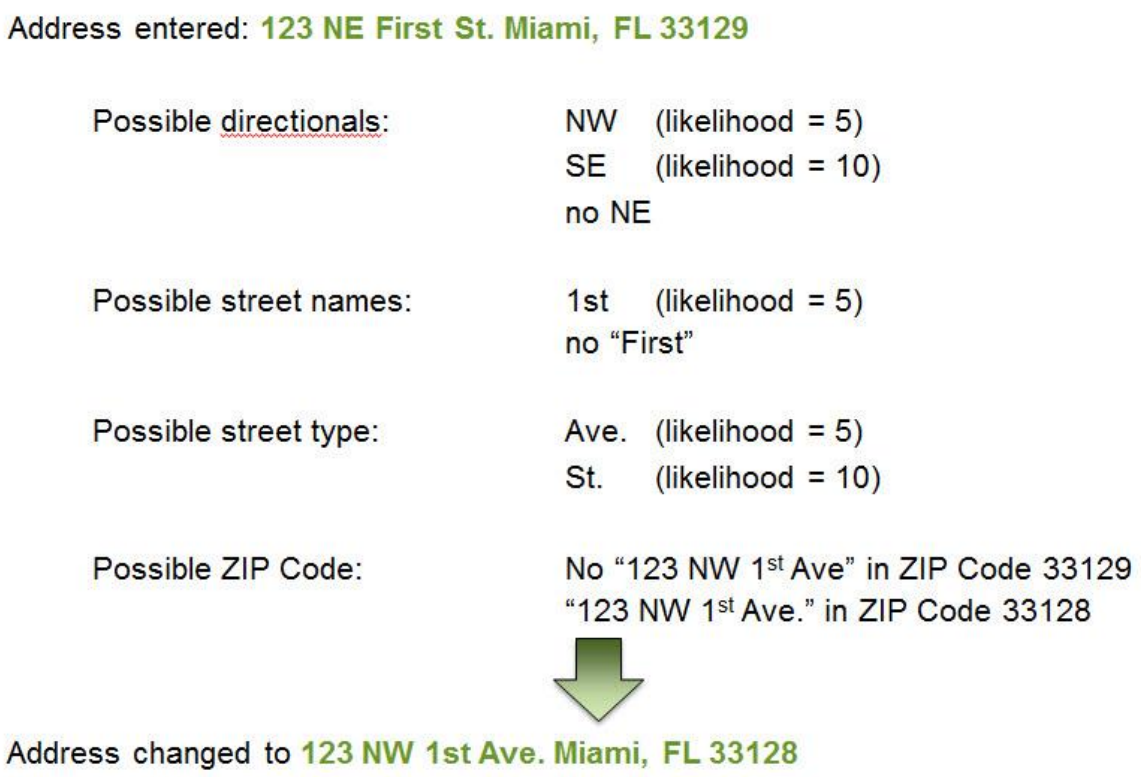
Address Standardization and Matching

In RMS products, users can change the system settings to either allow or disallow address changes. Allowing address changes will help the software compensate for human error introduced into the data.

Specifically, the geocoder uses a ranking system to figure out which elements of an address are the most likely to be correct. These rankings can be anywhere from 1 to 10, where 1 is the most likely to be correct and 10 is the least likely to be correct. The geocoder parses out all of the address elements, such as the street number, street name, street type (e.g. avenue or boulevard), and “directionals” (e.g. NW or SE). It matches each of these elements to the address database and finds the most likely matches by keeping the match with the lowest score.

For example, assume that an address is entered as “123 NE First St”, as shown in Figure 14. When the elements are parsed out, we find that there is no match for NE, and that NW has a likelihood of 5 and SE has a likelihood of 10. Furthermore, there is no match for Street. We only find a match on Avenue (likelihood of 5). As a result, the address is updated to “123 NW 1st Ave”.

FIGURE 14: Address Standardization and Matching Process

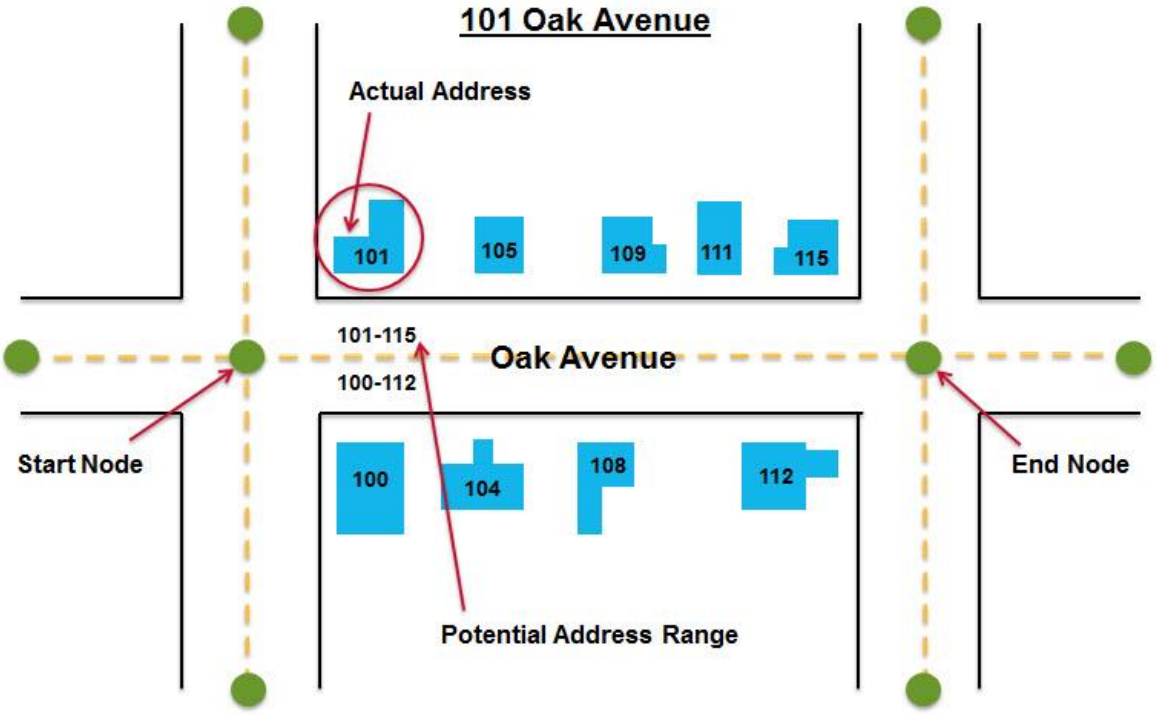


Street Address Interpolation

Once the correct address has been determined through the address standardization and matching processes, it can be assigned a latitude longitude coordinate through interpolation. Interpolation is the process of determining where exactly a location lies along a given street. The geocoder contains street segments with address ranges assigned to each side of the street. Typically, odd-numbered addresses are one side of the street segment and even-numbered address on the other. These street segments connect at each end point by a node. During the interpolation process, the address is located on the street segment using the percentage that the address is offset within the address range. For example, if the address “50 2nd Ave” is on a street segment with nodes of 0 and 100, it is positioned halfway along the

segment on the even-numbered side of the street. In the U.S. only, the geocoder also offsets the address from the street centerline so that it is not positioned in the middle of the street.

FIGURE 15: Example of Address Interpolation



Confidence in interpolation accuracy can vary by line of business. It usually works best in residential areas where house numbering and parcel size are relatively uniform. Heavy industrial or large commercial locations have the highest uncertainty because of the irregularity of the buildings relative to the street. Building level geocoding, such as that obtained using the Sanborn data, does not require interpolation because street segments are not used. Instead, building footprints are used to allow for a precise location placement of the specific building.

Building Level Geocoding for U.S.

A geocoding challenge in dense urban areas, especially for accumulation analysis, is the clustering of distinct addresses in a single building. The following figure shows the footprints of various buildings in downtown Chicago. The green crosses represent street level interpolated address locations. The location of some of these points relative to the building footprint shows that separate addresses are actually associated with the same building, which would not be captured using the traditional street level geocoding described previously. The black stars represent those locations geocoded to a building level.

FIGURE 16: Buildings Showing Street Level vs. Building Level Geocoding

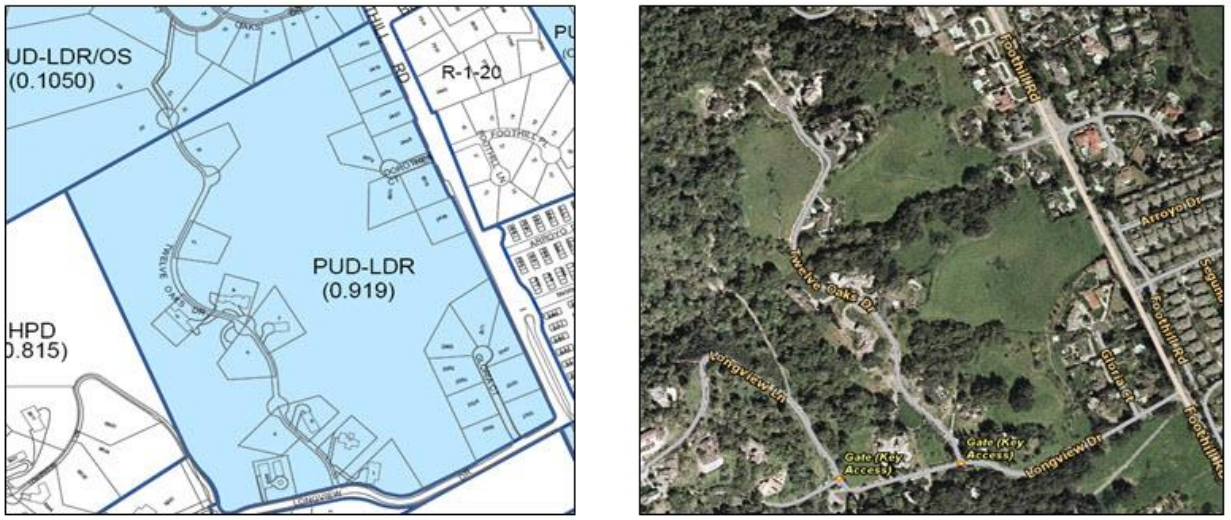
To address this challenge, RMS uses the High Resolution Building (HRB) data from the RMS ExposureSource Database (EDSB) that links multiple addresses to the same building footprint. When a building-level match is found, the geocoder will return a building ID as well as other values such as latitude and longitude. This BuildingID is stored in the EDM.address table and distinguishes locations geocoding to a building level from those with a street address match. Because building level geocoding uses footprints instead of street segments, no interpolation step is performed. Furthermore, attributes such as building height, number of stories, construction details, and occupancy type are stored for each footprint using this building ID. If these attributes were unavailable at data entry, they can be filled in once a building level geocode match is achieved.

Parcel Level Geocoding (U.S. and Australia Only)

Another geocoding challenge, particularly in residential and industrial areas, is the irregular distribution of exposure along a single block. The concept of street-address interpolation is based on the expectation that a block's address range represents a relatively uniform progression of properties from beginning to end, especially along a continuum from 00 to 99. In cases where this assumption does not reflect reality, geocoding results will be inaccurate.

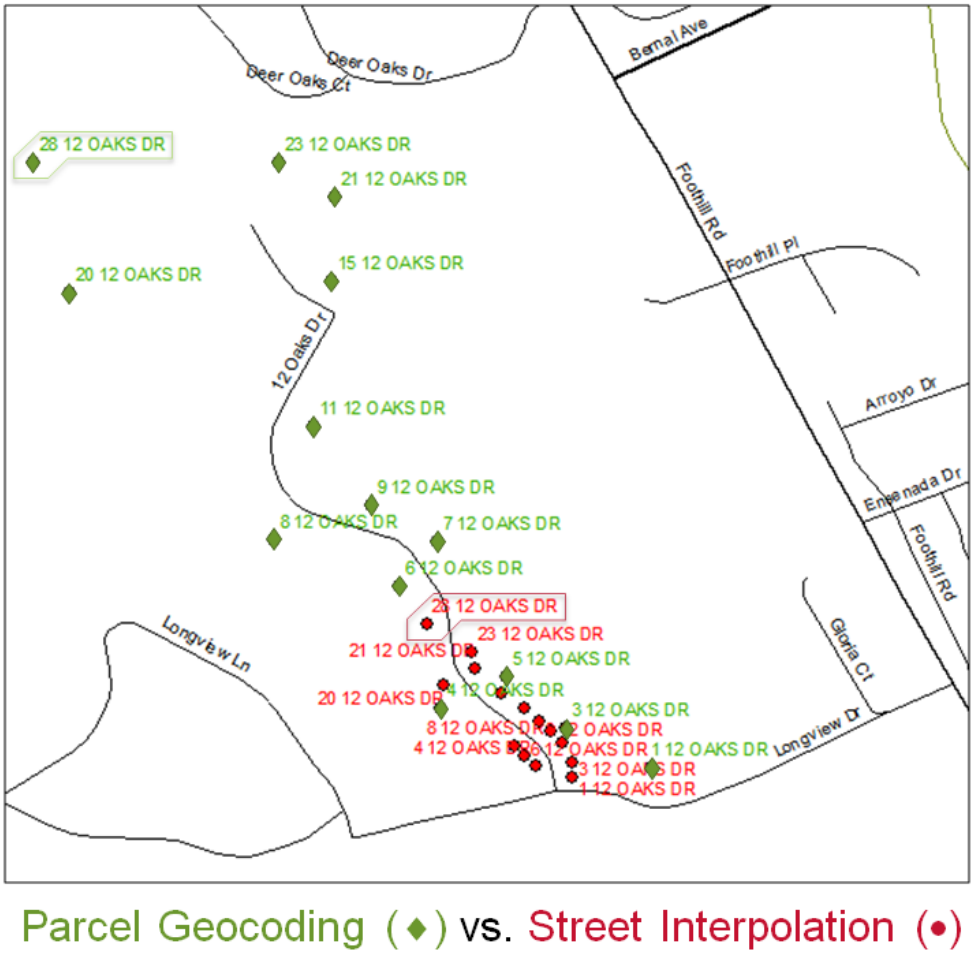
The figure below shows an aerial map compared to a parcel map, highlighting the distribution of individual parcel boundaries along a single block in Pleasanton, California.

FIGURE 17: Parcel Boundaries in Pleasanton, California



In this example in Figure 17, the geocoder assumes the addresses on this block range from 00 to 99, but instead they range only to 25. During the geocoding process, street interpolation places them all in the first quarter of the block. In contrast, parcel-level geocoding places them at the centroids of the actual parcel polygons. In Figure 18 on the following page, note how the mismatch becomes greater as the street numbers increase, with the biggest discrepancy for 2812 Oaks Drive (outlined on the map), the house at the end of the street.

FIGURE 18: Parcel versus Street Level Geocoding Results in Pleasanton, California



The Centrus Parcels Data Set is available for geocoding within RMS applications for an additional licensing fee. Centrus Parcels are based on parcel polygons acquired from counties across the United States. With this information, the Parcel Data Set enables point-in-polygon parcel geocoding for many areas of the United States. Parcel-level geocoding provides better spatial accuracy than any street interpolation data set, and is more spatially consistent over time. These factors combine to support better analyses and produce smaller changes year to year.

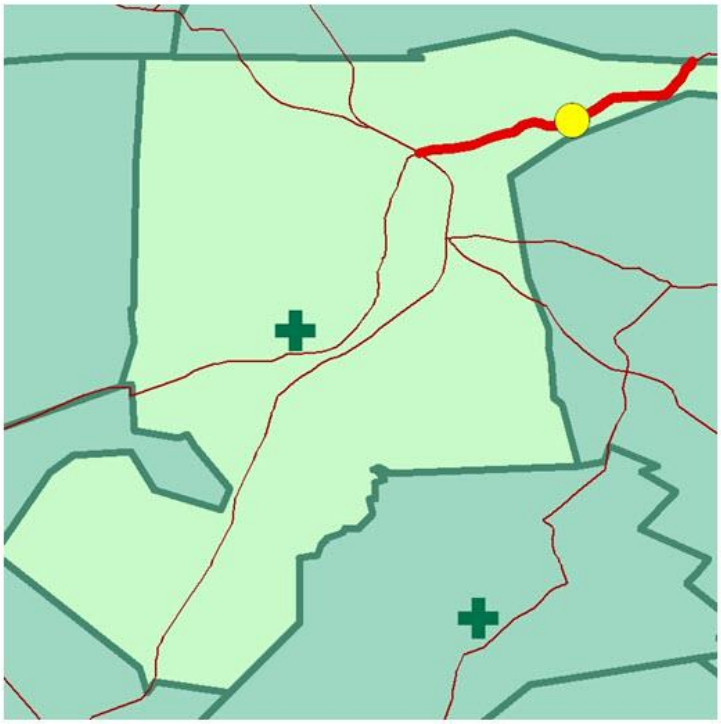
Street Name Matching Outside the U.S.

For locations outside the U.S., it is common for accurate address numbers to be unavailable for some location addresses, or for street segments to not include address ranges. Street-name matching is available through the GLM as one means of addressing this challenge. Street name resolution falls between the street address and postcode levels and provides a midpoint location of the segment for the street name entered. If a postcode is also present, the geocoder matches on the street name and returns the latitude-longitude midpoint of the street segment falling

within the postcode. The result is not as accurate as a street address match, but more accurate than a postcode match.

The map in Figure 19 below shows a number of street segments crossing through a postal code in Belgium. The postal code centroids are indicated by the + symbol. Street segments can be found by name and a latitude/longitude coordinate pair is assigned representing the centroid of that segment. Note that in this map, the street segment is in bold red and the street centroid is indicated by the yellow dot. Being able to geocode at this level provides a much better level of accuracy than geocoding to the postal code centroid, which was the best available option prior to the GLM.

FIGURE 19: Street Name Geocoding in Belgium



E-LEARNING INTERACTION 3: Street Name Matching: How it Works

To review an interactive example of the street name matching process, go to the Geocoding and Hazard Retrieval course in the CCRA portal of Owl and select Interaction 3: Street Name Matching.

Geocoding Changes Impact Loss Results

It is important to note that when underlying geocoding databases change, it can affect geocoding results and model loss results. The levels of change in geocoding and in loss are not always correlated. Only if the change in geocoding resolution or positional accuracy results in a change of hazard or vulnerability will the loss results change. This is based on the specific characteristics of the site and varies with each portfolio.

When geocoding changes result in modeled loss changes, most of the variation in loss results is a result of improved geocoding resolutions. The locations that geocoded at the street address resolution retrieved their hazard data from the high resolution VRG, instead of the more aggregate postal code. Such a change introduces the potential for very different location level loss results, and a corresponding potential to influence losses at the portfolio level.

Unit 2: Closing Key Concepts

This unit focused on the technology that underlies the geocoding process, specifically software functionality and high resolution geocoding match levels. In addition, the chapter emphasized address standardization and matching, street-level interpolation, geocoding both in and outside of the United States, and filling in the geocoding framework.

Finally, it is important to remember that when underlying geocoding databases change, model loss results can change as well. Understanding the changes to the geocoding software is the key to understanding the downstream impacts on results.

Unit 3: Address Systems and Case Studies

This unit will cover how to maximize success with geocoding technology, specifically through the quality of location data and troubleshooting address errors. In addition, we look at Japan as a case study in order to illustrate the importance of being familiar with local address systems and model requirements.

Learning Objectives:

- Develop an understanding of geocoding with diverse address systems.
- Match address elements with specific modeling requirements.
- Demonstrate how to make sense of unfamiliar address systems.

Understanding Local Address Systems

RMS geocoding technologies are designed to accommodate the address systems of countries worldwide. Most countries have their own unique and culturally accepted addressing methodology. Some countries such as Sweden, Australia and the U.S. have a relatively short, two- or three-line address, while others require much more information, such as Japan and the U.K. Some countries use a sequential numbering system for buildings on a street that go from one end of the street to the other. In other countries, buildings on a street may be numbered chronologically by the order in which they were built.

Some of the potential issues that add to the complexity of international address systems include translation from the local language and transliteration of the original alphabet. The latter may include special characters that can be written several ways. Abbreviations may not always be apparent, and the systems for ordering or numbering may not be intuitive to the uninitiated.

In order to understand addresses in different countries, it is important to first identify each of the address elements for that country, and then recognize what is most important for modeling. Because countries have different address formats with unique features, the geocoder must be able to recognize several levels of resolution and record match levels.

FIGURE 20: Examples of Diverse Address Elements

Definition	France	Germany	Australia
Addressee Name	Monsieur Yves Lefort	Rudolf Schneider	Mr. Michael Hughes
Building Name	Escalier No. 1	Bingerbruck	
Street Address	12 Boulevard de la Cordiere	Kreuzbacher Str. 48	57 Alexandra Ave
Postcode and City Name	13007 Marseille	55411 Bingen	Canterbury VIC 3126
Extra Identifier	Cedex 1		

Looking at the table above, in France the geocoder recognizes street address (including street name), postal code, city, département (the French equivalent of a county), and CRESTA. When looking at the address, the street address and name, city, and postal code look similar to those in the U.S., where the street number precedes the street name. The département is the first two digits of the postal code (in this case, 13) in metropolitan areas, and you may also see a Cedex listed, which is a separate address coding used only in France.

By contrast, in Germany the street number follows the street name. Similar to France, however, the postal code is listed first and then the city name.

Australia uses the same address format as the U.S. and Canada. It has four-digit numeric postal codes, plus six states and three territories which are always abbreviated in caps (e.g. VIC = Victoria). The first digit of the postal code indicates the state or territory. The last three digits represent the delivery area.

For anyone managing an international portfolio, understanding these address systems is vital. Once the role of each international address element is understood, you will be better prepared to identify those that are most useful in an RMS models.

The RMS document *Geocoding and Analysis Levels for RiskLink® 18.0* lists all of the levels of address geography that are useful for modeling.

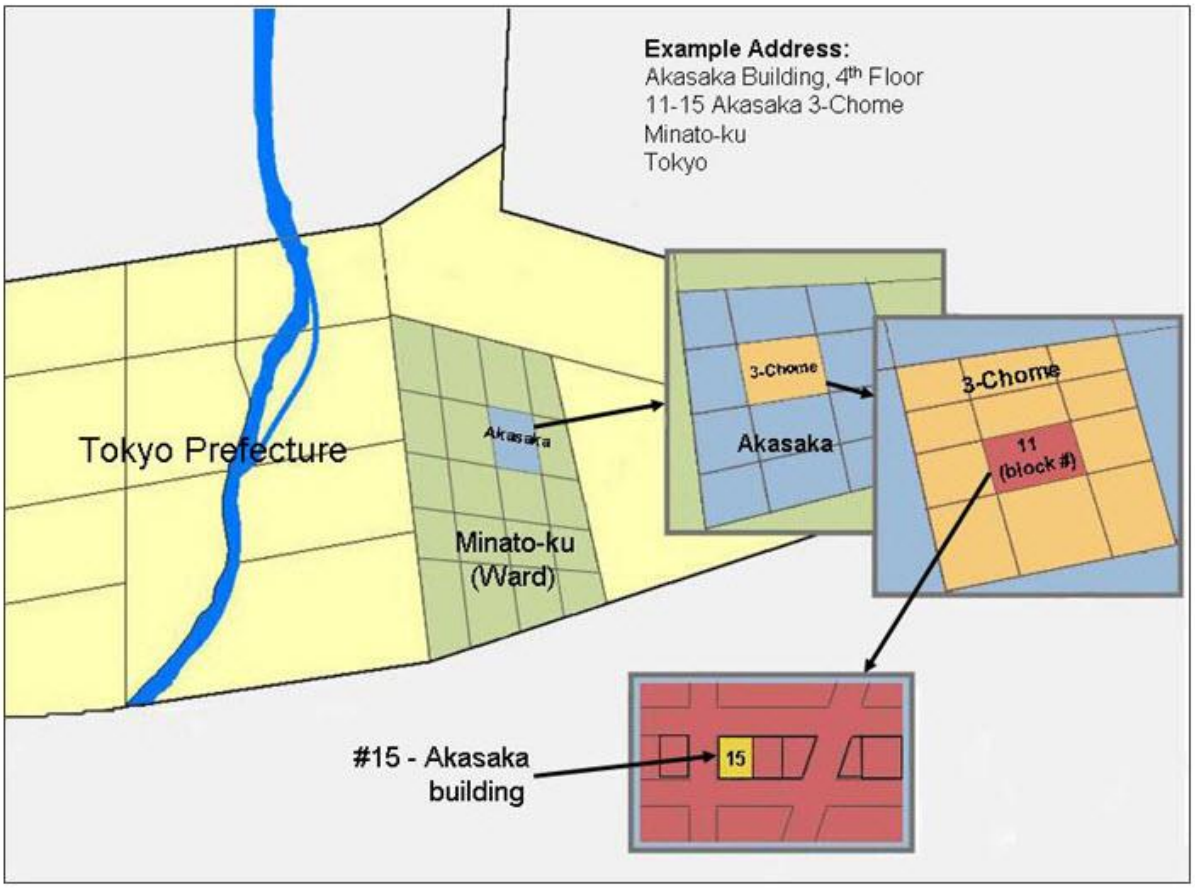
Making Sense of Japan

The Japanese address system is one of the most complex in the world and is useful in illustrating the diversity in these schemes. Japan is just one country that does not use one of the linear addressing approaches typical in most Western nations; India is another. Japan instead has a concentric address system and only the largest streets have names.

Instead of a linear methodology with streets and address numbers that start at one end of the street and end at another, Japan is based on geographic areas that are divided into smaller and smaller pieces. Streets have addresses, but they are not sequentially numbered from one end of the street to the other.

Japanese addresses are designated for individual lots, and address numbers are given in chronological order by the date that the property was developed, not by what position it occupies on the block. The lower numbers are assigned to the first buildings constructed and progress to higher street numbers for newer buildings. The lot is just the smallest element in a huge system of geographic units that start large with prefectures and are subdivided into successively smaller units like the Ku (ward) and Chome (district). The specific format of an address in Japan can vary, depending on whether it is in a rural or urban area.

FIGURE 21: Making Sense of Japan

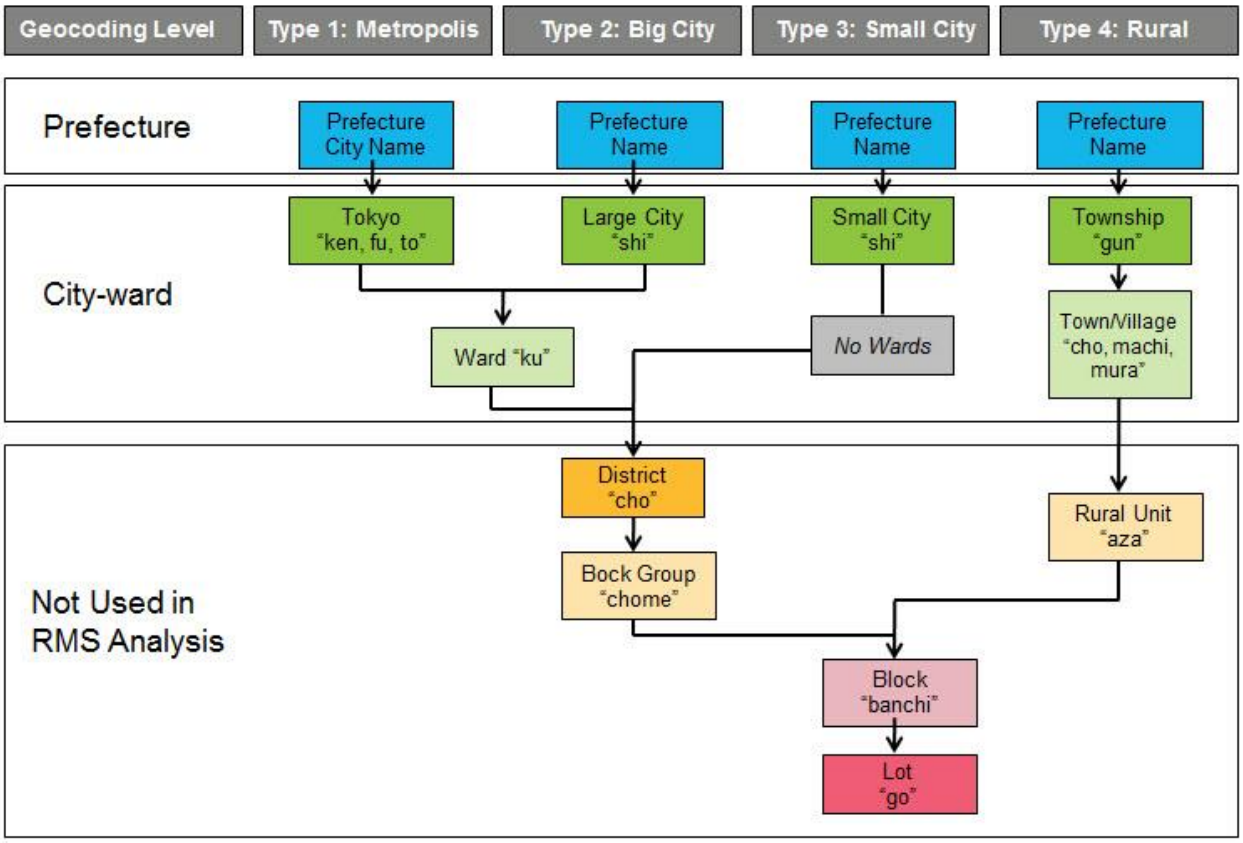


The Japanese address system is atypical and is provided as an example to illustrate the key point: it is important to have at least a basic understanding of all elements in an address system in order to identify the address elements that are used in and thus are important for a catastrophe model analyses. The three most important elements to be able to identify when using RMS models are prefectures, city/wards, and postal codes.

The chart in Figure 22 may be a useful reference to identify which address elements are used by the RMS Japan models. The three most important elements to be able to identify in a Japanese address are prefectures, city/wards, and postal codes.

- Prefectures: There are 47 prefectures in Japan. Each prefecture has its own unique name, without extensions. The exceptions are if one of the top five cities in Japan resides within the prefecture, in which case the prefecture and city names are the same. These top five cities are Tokyo, Yokohama, Osaka, Kawasaki, and Kobe.
- City/Wards are identified with the prefectures "-ku" or "-shi." For example, "Chiyoda-ku" is the ward of Chiyoda within metropolitan Tokyo.
- Postal codes are easier to understand for westerners because they contain easily recognizable Arabic numerals. However, they can be easily confused with Sonpo codes in Japan, which can be used to identify city/wards, so it is important to properly identify these numeric values as one or the other.

FIGURE 22: Japan Address Elements



E-LEARNING INTERACTION 4: Making Sense of Japan

To further see how the Japanese address system works, go to the Geocoding and Hazard Retrieval course in the CCRA portal of Owl and select Interaction 3: Making Sense of Japan. The purpose of this interaction is to highlight an example of a complex address system. Specific details regarding the Japanese address system are not included in the CCRA exam.

Troubleshooting Address Errors

Errors can make their way into location databases no matter which country you are working with. Most address errors occur during the process of data transfer and translation. The process of collecting data can involve policyholders, agents, primary insurers, intermediaries, and reinsurers. Each transfer or format change has the potential to alter the modelled position of a location.

Some of the more basic errors include typographical mistakes or errors when data is converted to a different format to accommodate different computing systems. When dealing with multi-national data, it is easy to introduce errors when translating between languages. Also, if you are unfamiliar with the local address systems, it can be difficult to recognize where data should go or recognize existing address errors.

It is important to note that spelling is one of the biggest challenges in geocoding international locations. Different languages tend to use different transliteration models, not all of which are recognized by the RMS geocoder. For example, English transliteration of Japanese addresses may use the spelling “Ibaraki” while German uses “Ibaragi.” Another example is Gotemba vs. Gotenba. Germany tends to favour the “-mba” rather than the “-nba,” but only the latter (Gotenba) is recognized by the RMS geocoder.

RMS has published resources that will help you understand which elements you need to pick out of an address that will get you the best geocoding and modeling results. When working with a submission for a country with which you are not familiar, it is good practice to reference these tools and resources.

Unit 3: Closing Key Concepts

Countries throughout the world have different address formats with unique features. The key to understanding these diverse address systems is to identify the address elements for each specific country, and then determine which is the most important for modeling.

In some countries, there is some address information that may be captured and stored for querying but may not be relevant to geocoding within the models. This may be because other address elements are sufficient on their own, the data are not available within the geocoding engine, or the model is built to a lower resolution.

Unit 4: Hazard Data and Retrieval

This unit presents information on hazard data retrieval and the impact on the modeling process. It examines the relationship of model data flow, resolution vs. coverage in hazard data, and the evolution of hazard data over time. This unit covers earthquake modeling and site hazards, including ground-shaking and liquefaction data, landslide data, hazard aggregation, and exposure weighting. It will also cover hurricane modeling and event hazards, including surface roughness and wind direction, topography and wind direction, the evolution of wind fields over time, and the relevance of distance to the coastline.

Learning Objectives:

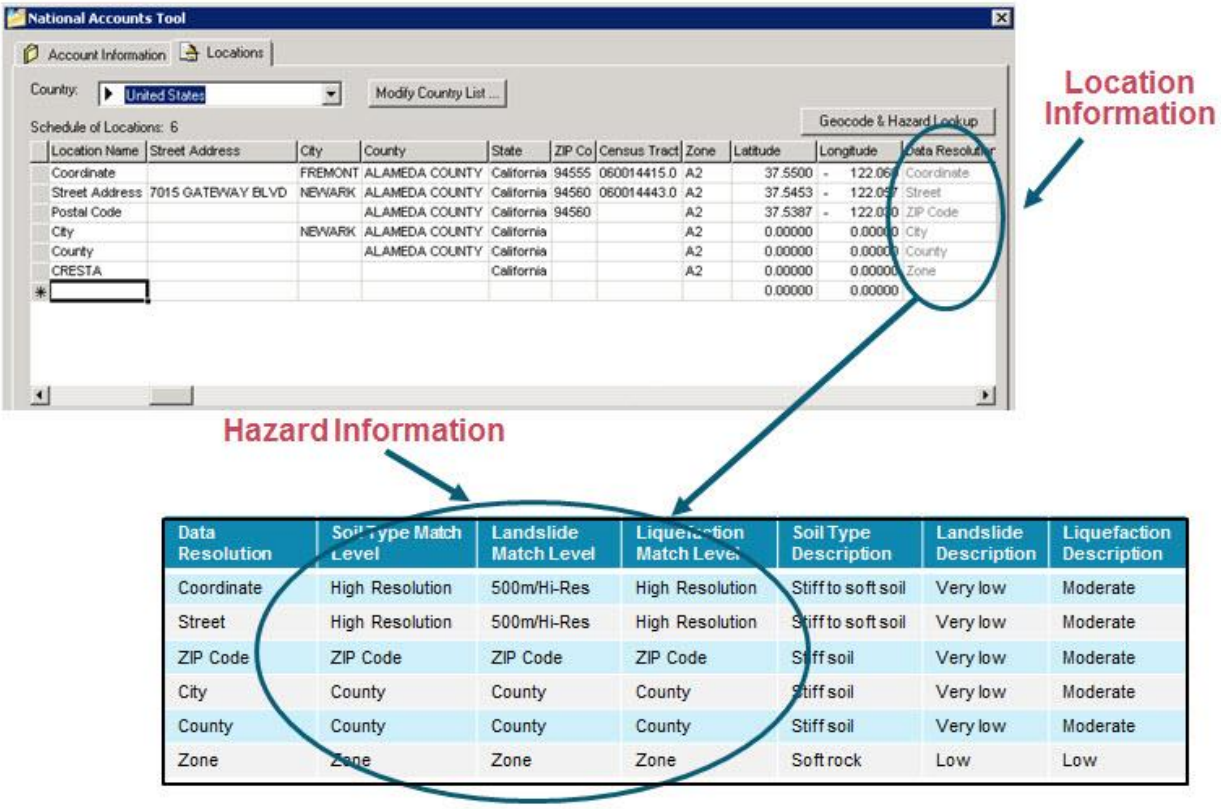
- Describe the tradeoffs between hazard data resolution and coverage.
- Describe the influence of specific hazard variables on risk.
- Understand the relationship between exposure data quality and hazard data resolution over time.
- Explain the difference between site hazards and event hazards.
- Explain the role of spatial accuracy in hazard retrieval.
- Understand the process and benefits of weighting exposure values.

Geocoding and Hazard Site Resolution

Following the geocoding process, the next function is the hazard retrieval. Hazard data detail varies by region and peril, from high resolution maps (soil, flood hazard, geologic, etc.) to data aggregated to coarser units, such as postal codes, cities and counties.

The hazard retrieval follows from the geocoding match. In other words, **the hazard data resolution retrieved can be as good as, but never better than, the geocoding level.** Even if detailed hazard data are available, the resolution used in modeling is dependent on the quality of the geocoding match. If a location geocodes to a postal code level, then the hazard retrieval will also be done at the postal code level, even if more detailed hazard information may be available.

FIGURE 23: Hazard Match Follows from Geocoding Match



Detailed hazard data are typically not available over large areas. For example, in the U.S., nationwide soil coverage is available, but the highest detail data are available only for the most earthquake exposed areas. There are multiple reasons for this. Source maps for high-resolution data are not always available or suitable for use in risk models, for one, and there is also a need to balance greater spatial coverage versus greater detail in highly-exposed regions (Figure 24). RMS continues to work to increase the level of coverage and detail in all of the world's earthquake exposed areas. It is important to collect good quality address data now, so that as finer hazard resolution becomes available, locations geocoded at the high resolution level can access the new data immediately.

FIGURE 24: Resolution versus Coverage in Hazard Data

The Variable Resolution Grid (VRG)

RMS models have historically used either high resolution or postal code level data. But a gap exists (spatially) between these. Postal code boundaries can change, sometimes frequently. Postal codes are driven by mail delivery needs and are based on population. Furthermore, postal code areas are smaller in urbanized areas with higher exposure (meaning insured values), but they do not represent exposure to a given peril, and they do not take hazard into account at all.

RMS developed the Variable Grid Resolution (VRG) in response to two needs: (1) the need for better data over larger areas, and (2) the need for stable hazard data, i.e. data that would not change at the direction of postal services. The VRG is a grid system where the size of each cell varies depending on the level of risk and exposure. In areas of either high risk or high exposure (or both) the grids could be as small as 50x50m (e.g. Belgium flood). In areas where the risk and/or exposure are less, the grid size could be as large as 100x100km (e.g. southwestern Australia). The VRG grid boundaries are designed specifically to meet the needs of the insurance industry for catastrophe modeling.

VRG takes the highest resolution of hazard data available and aggregates it to the grid level. Data is still available at lower levels such as the postal code level, so the model can retrieve hazard data when only postal code data is available.

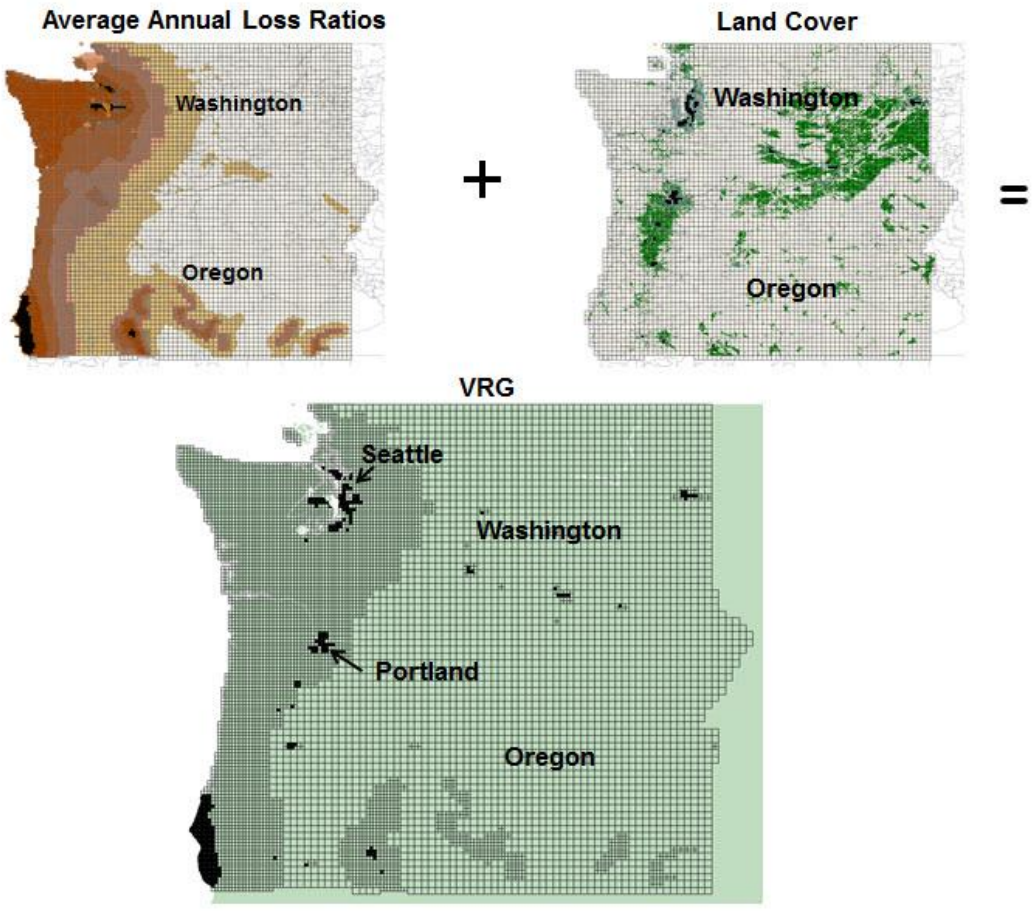
The VRG design accounts for the level of risk, considering both hazard and exposure. The VRG infrastructure is designed to focus model detail on the areas where it can have the greatest impact.

Figure 25 shows maps of Washington and Oregon, states in the northwestern U.S. The left map shows gradations in earthquake hazard. The gradations tell us where the probabilistic ground motions are highest, which is clearly along the western third of the states in this example. Furthermore, there is pronounced hazard in the Seattle/Tacoma areas and near several faults in southern Oregon.

The map on the right is a land cover map, shaded in colors representing different types of urban and rural use. These data can be used to identify areas of denser exposure. Most of these areas are in the Seattle/Tacoma metro area, the Willamette Valley of Oregon, and in the East, near Spokane.

RMS combines these maps to produce a VRG that considers both sets of data and features the smallest cells in the areas of greatest exposure and hazard. Notice that in Portland and Seattle the cells appear black – that is because they are so small in these areas they cannot be distinguished without a more detailed zoom level.

FIGURE 25: Combining Hazard and Exposure Grids into One VRG for Earthquake

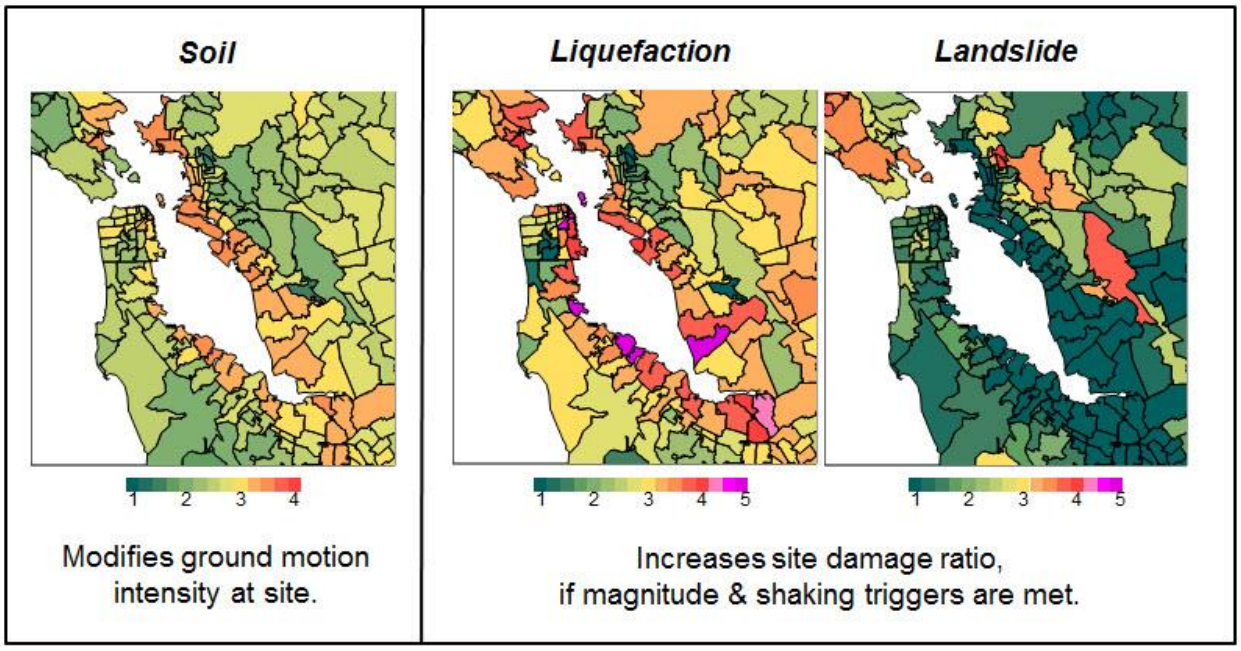


The VRG design is different for every peril; in some models, the grid changes by event to capture the individual variations in hazard. The same principles are applied, however, in that the detail is always focused on the areas with the greatest risk. For hurricane, the VRG provides the most detail where hurricane hazard and insured exposure are both greatest. The result in Florida is that the smallest cells are located near the coast. Regardless of the peril, the VRG helps to provide appropriate resolution risk differentiation while improving model performance.

Earthquake Modeling and Site Hazards

In working with an earthquake model, certain hazard variables are characteristics specific to the site, rather than a given event, and are retrieved independently. There are three variables linked to the local geology: soil, landslide, and liquefaction. Figure 26 shows hazard data by postal code for the San Francisco Bay Area in California. Each of these captures an aspect of how a given site may respond to the energy released from an earthquake. “Soil” in this context is an index based on the physical properties of surficial geology that influence the local ground motion at a site; rock or stiff soil can dampen the severity of shaking whereas soft soil tends to amplify it. Liquefaction and landslide provide an index of the susceptibility of the local materials to these types of ground failure, which can increase the damage at a site beyond that caused by ground shaking.

FIGURE 26: Earthquake Site Hazards



This data used in Figure 26, shown at the ZIP Code level, is in a digital format and is available for use in RMS models. But where does RMS get these data? And how much confidence can we have in their accuracy? Some of the RMS hazard data have their source in paper maps, while some of the data we have acquired in the more recent years are available digitally.

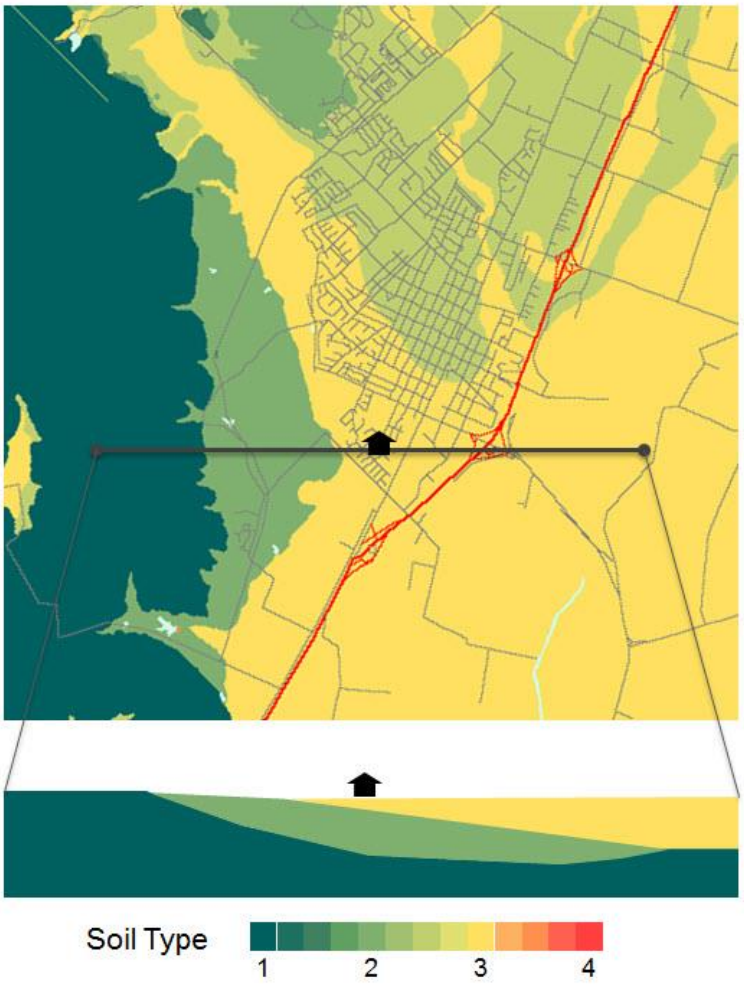
RMS engineers have close relationships with public and private geologic organizations in every country where we have earthquake models. Our engineers acquire these data as digital databases or in their original paper format. In the case of paper maps, the hard copy versions are digitized in our data production facility. Once all of the data are in digital form, our engineers perform a classification of the map content. The individual geologic units are boiled down on the basis of their physical properties to a smaller number of standard classes that are meaningful for our models.

For example, the soil classes reflect the observed correlations between shear wave velocity, a material property of soil, within the top 30m (100 ft.) or so of the surface and ground motion amplification. When earthquake energy hits soft soils, the reduction in the speed at which the ground motion can pass through the material can result in the amplitude of the ground motion increasing, particularly when there is a sharp contrast in shear wave velocity with underlying materials. Rock and very stiff, consolidated soils, on the other hand, is less likely to increase the shaking at a given site. While there are multiple factors that can influence the level of ground motion at a site in an earthquake event, the soil provides an indication of the likely relative level of amplification.

RMS soil classifications have been refined over time and have evolved from integer values in early models to a continuous decimal scale. The decimal implementation recognizes two aspects of the data that are relevant for modeling. One challenge is that the boundaries between soil types are usually transitional rather than distinct and there is inherent uncertainty in classification. Assigning intermediate classes allows these transitions to be applied.

Observe Figure 27, which shows a geologic map and cross-section. From the map of the surface materials, it is usually not possible to know what's happening below a site without additional data from boreholes or subsurface imaging tools. Given that the thickness of the geologic materials can influence their shaking response, it can be difficult to truly know where one soil ends, and another begins. This is one of the many contributors to uncertainty in model results.

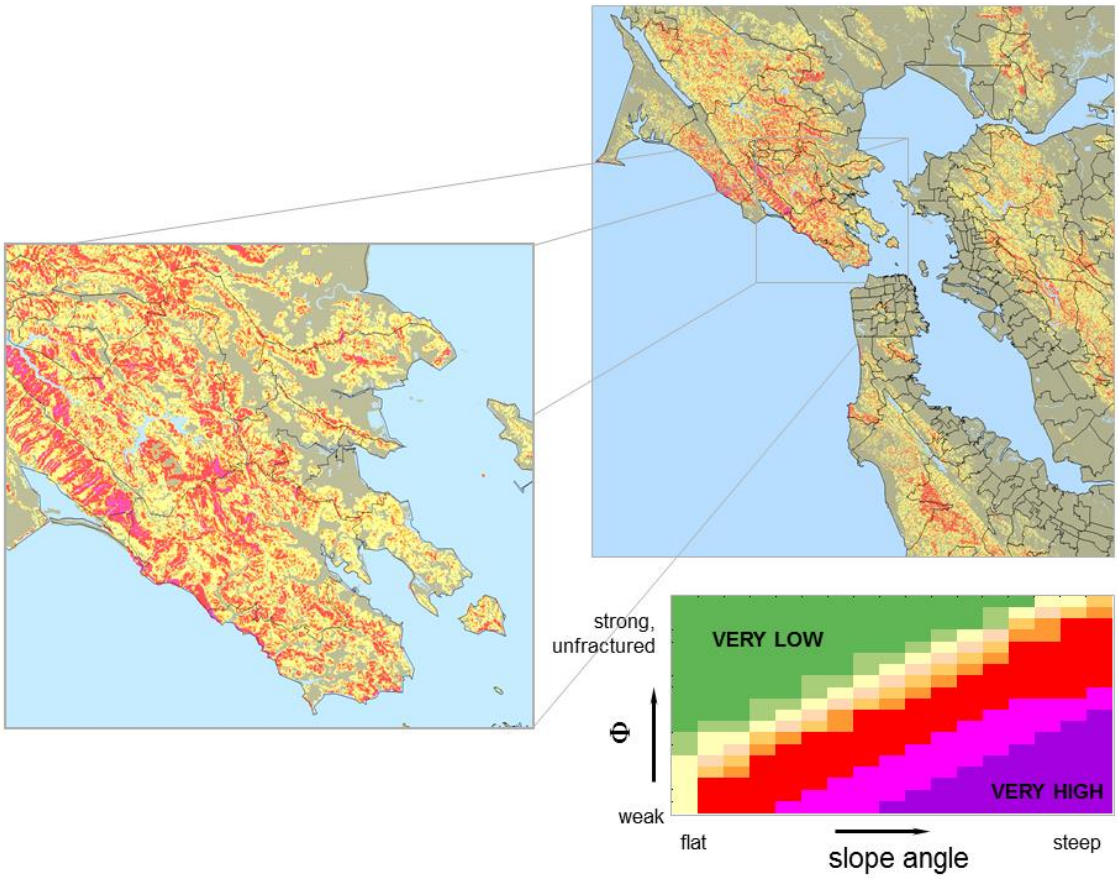
FIGURE 27: Soil Map Cross-Section



A second is in the delivery of data at aggregate resolutions. A given postal code may contain multiple soil classes, but if a location can only be geocoded to the postal code resolution, the attributed value will be some kind of weighted average. This point is discussed further on in this unit.

Sources of landslide susceptibility data are harder to come by and these layers require additional modeling steps in their development that illustrate the impact of data resolution. One approach to developing landslide susceptibilities uses a combination of geologic and topographic data, calibrated when possible against inventory maps of historical landslides. Applied by the California Geologic Survey (CGS) for selected urban areas, the classification uses material properties such as cohesion and the coefficient of friction (represented by ϕ in the graph below) indicates how well a given soil or rock type resists failure. Slope is calculated from digital elevation model (DEM). Each cell within the area is then classified using a matrix that combines the slope with its resistance to failure (Figure 28). The matrix assignments can vary depending on whether the regional conditions are wet or dry.

FIGURE 28: Integration of Multiple Data Sources for Landslide Susceptibility



It is important to keep in mind that because the development process incorporates a *general set of assumptions* to develop data that help us better understand the risk of landslide, there is inherent uncertainty within the data. Site-specific studies performed by geotechnical engineers provide a more accurate evaluation of individual risks, as they can account for additional factors that are tailored to the location. The time and cost involved in these usually make them impractical for all but high value properties with significant hazard.

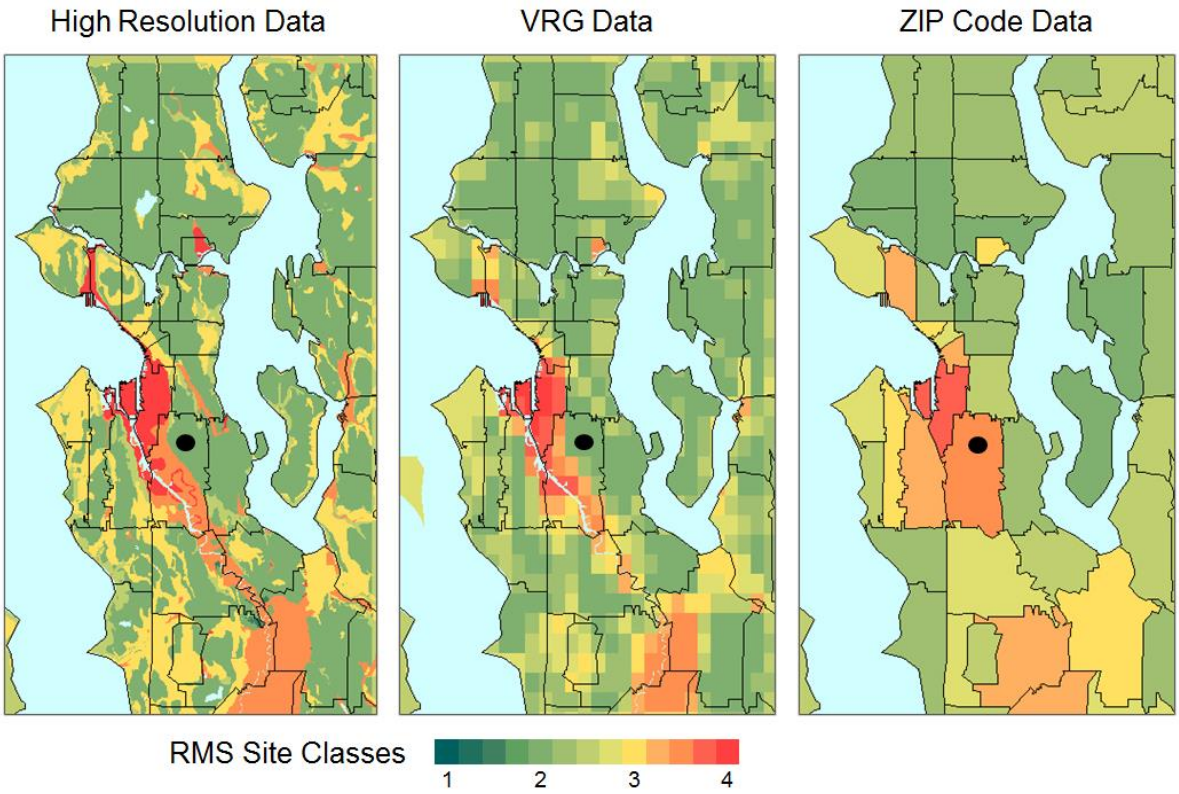
Managing Data Precision: Hazard Aggregation

Aggregate data are employed in all hazard data and models. These aggregated hazard values result in use of the continuous decimal scale, averaging a potentially large number of values from throughout a broad area. It rarely shows abrupt differences over short distances like high resolution soil data does, which is another reason that high resolution data are preferred over coarser aggregate resolutions such as postal code or county.

Figure 29 illustrates the change in soil values between high resolution, VRG aggregate, and ZIP Code aggregate data for the Seattle, Washington region in the U.S. The black lines are ZIP

Code boundaries and the black dot represents a hypothetical location. Soft or poor soils are in red and orange, harder/better soils are in yellow and green.

FIGURE 29: Aggregation of Soil Types in Seattle, Washington

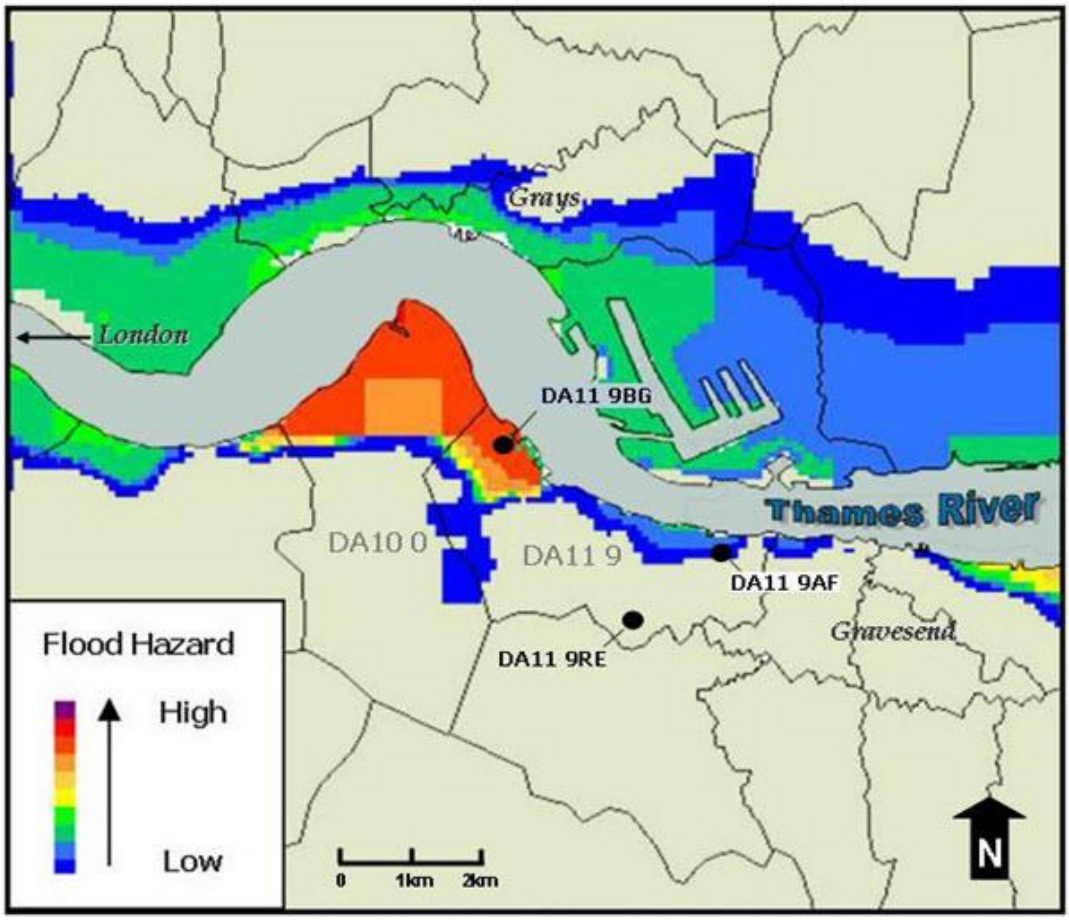


The high resolution data show that the location would return a green soil lookup, which represents harder/better soil. In the center panel, in which the high resolution data have been aggregated to a VRG cells, you will notice the blocky appearance and transitional colors of the map. In this case, the location would still return a green soil lookup, just as it did for the high resolution soil map. The ZIP Code data are aggregated even further to larger areas. The value in each ZIP Code represents the weighted average soil type, but there could be significant variation within the area. Because of the aggregation technique used (more on this later), the predominant soil type for the sample location using this map was determined to be orange, or not very good.

This illustrates the versatility of the VRG, compared with ZIP codes, as a geographical framework for organizing hazard data. It also illustrates the importance of good location data. If this location geocodes to a street address, it gets assigned a good soil class that dampens ground shaking and may reduce modeled loss. If only a ZIP Code is available, in this case the location is assigned a poorer soil class that will increase modeled loss. Note that older model versions only had high resolution or ZIP Code soil data available.

As another example, Figure 30 illustrates a section along the River Thames just east of London, U.K. The dark grey boundaries are postcode sectors, a geocoding resolution similar to U.S. ZIP Codes. Note Sector DA11 9 and the variation in flood hazard within the sector, especially along the river. This also happens to be in an area of high exposure. If only sector level hazard data were available, one uniform hazard value would apply to the entire sector, even though parts of the sector are not prone to flooding. There are three postcode unit points displayed (similar to U.S. nine-digit ZIP Code points). Geocoding to this level would allow the user to access the detail of the VRG level data. Additionally, some VRG cells show distinct variation in size. The smaller cells are indicative of higher exposure, and the larger cells are in areas that are less developed. The VRG helps us to capture risk more accurately in areas of high hazard variability.

FIGURE 30: Flood Hazard by VRG along the River Thames near London



The precision of data varies by country and peril. Even if high-resolution data are available, the resolution of the hazard information retrieved for each site is limited by the geocoding match level.

Hazard Aggregation and Exposure Weighting

Two commonly-used methods for hazard aggregation are area weighting and exposure weighting.

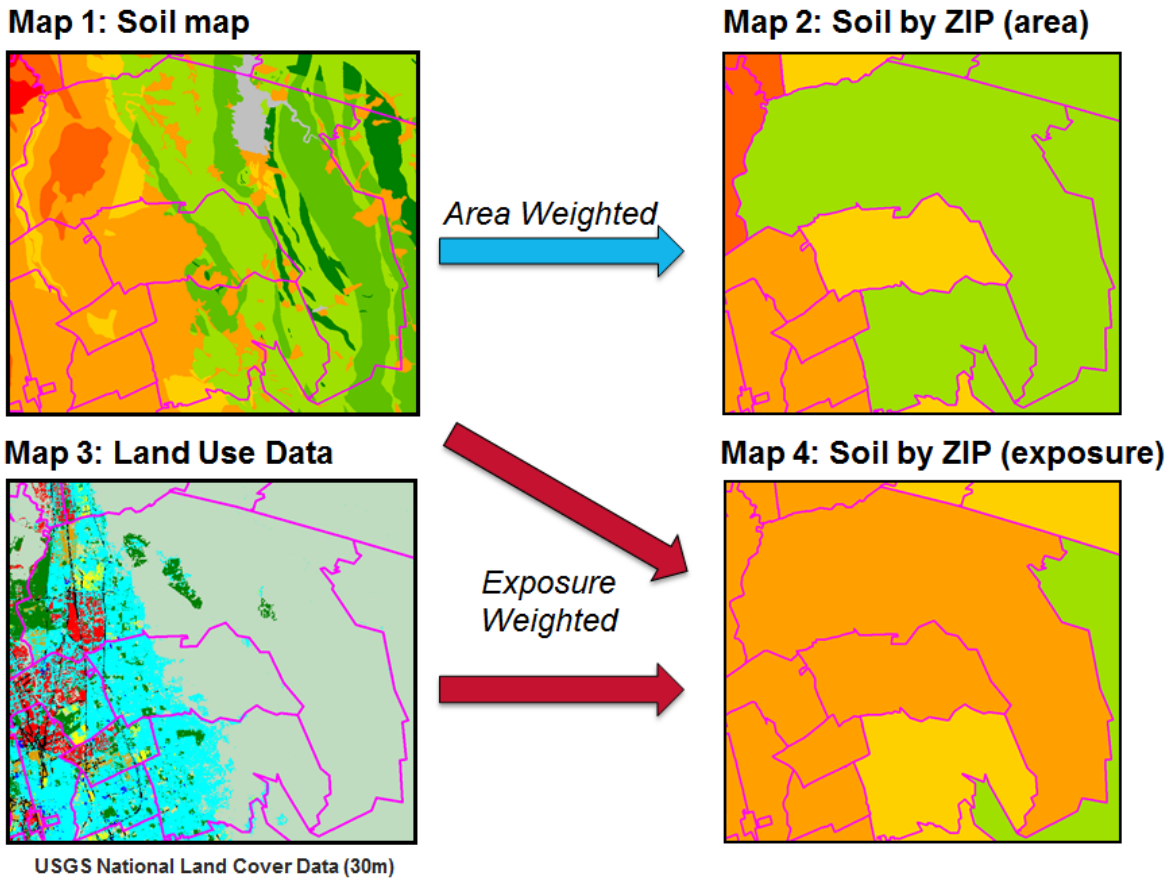
Figure 31 on the following page shows four maps. Map 1 shows the original hi-resolution, unweighted soil data with postal code boundaries in pink. Poor soils are in red, and the best soils (from an earthquake standpoint) are in dark green.

Map 2 shows data by ZIP Code that have been area weighted. This methodology assigns each soil type within the postal code a weight according to its proportion of the unit and then calculates an average. For example, if 40% of the ZIP Code was soil 1.0 and 60% soil 3.0, the result would be 2.2. You can see that for the large postal code in the center of the map, the soil type is light green, which represents good soil.

Map 3 shows land cover classes that were generated using satellite image data. The area of greatest development/highest exposure is to the west, shown in yellows and reds. The grey area in the eastern side of this ZIP Code is undeveloped. Note that this is where most of the better (green) soils are.

Map 4 shows the outcome of exposure weighting, the process of aggregation typically used by RMS. RMS takes the land use data and assigns each class a relative exposure weight. Urban cells have weights that are larger than suburban and are many times higher than forests or otherwise undeveloped land. Using GIS tools, these exposure weights are overlain on soil classes within the postcode and used to assign a higher significance to the areas where buildings are located. In the example of Map 4, the large ZIP Code in the center of the map is assigned an orange soil type (worse soil) because most of the development in this ZIP Code is in the west on the poorer soils.

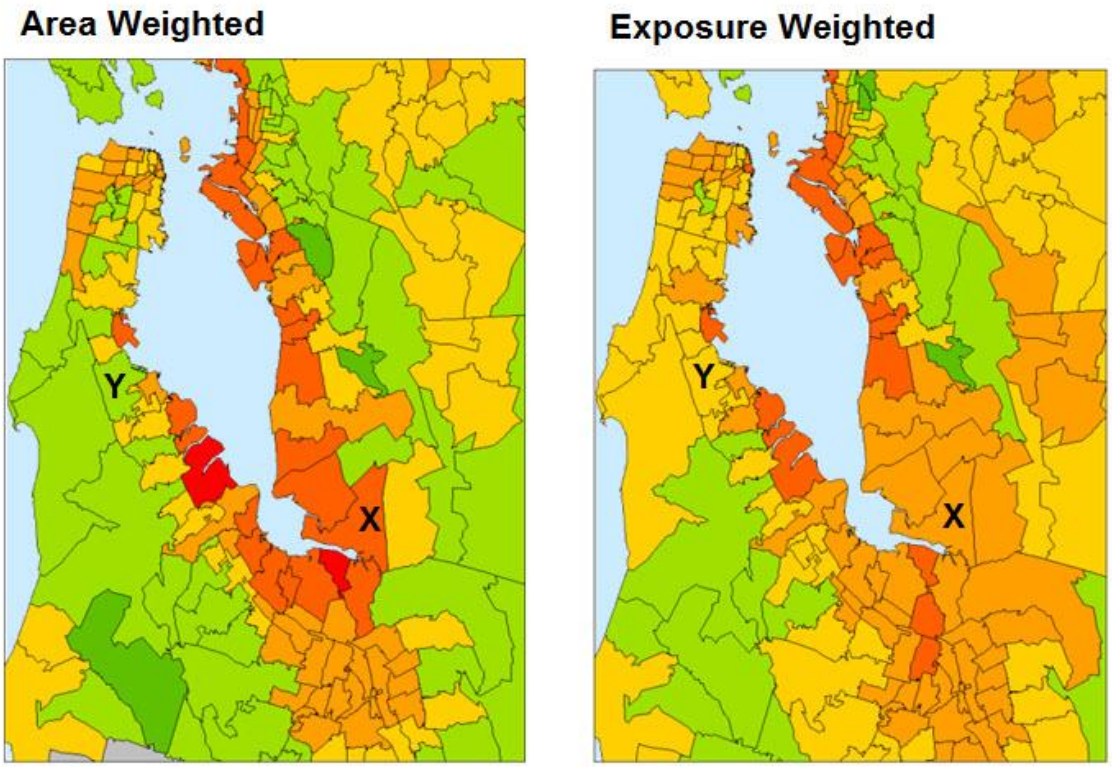
FIGURE 31: Exposure Weighting Examples for Soils



Another look at the two aggregation methods is show in Figure 32. These maps compare the two aggregation methods for soil hazard values in the San Francisco Bay area in California. Poor soils are in red and orange, and generally fringe the southern half of the San Francisco Bay.

Note that exposure weighted aggregation (map to the right) does not always cause the hazard assignments to shift to poorer values; the outcome varies by ZIP Code. For example, in ZIP Codes around the southern tip of the bay near the location marked X, exposure tends to be located on better soils within the ZIP Code boundaries. Exposure on the mid-peninsula near the location marked Y tends toward less favorable soil classes within the ZIP Code boundaries.

FIGURE 32: Regional Effects of Exposure Weighting



Soil-hazard values at the ZIP Code level in the San Francisco Bay Area, California

These examples illustrate some of the variations and assumptions that are embedded in use of an aggregate hazard resolution. It can be seen that the exposure weights assigned to the land use cells will influence the outcome. In fact, the area-weighting approach is actually identical to assigning all land use types the same weight. The final goal is to provide a value that represents the average condition for exposure within the aggregate unit, but any individual location within that unit could have a value that is very different.

Hurricane Modeling and Event Hazards

Earthquakes are directly related to movement along faults, identifiable physical features in the Earth's crust. Common practice is for events to be treated as either specific, fixed points or rupture areas, from which ground motion radiates outwards. These are usually treated as occurring instantaneously[1], given that the rupture and resultant shaking take place over minutes at most. As a result, we say that earthquakes are static events – not in the sense of any lack of motion, of course, but that where they occur can be modeled in fixed locations that have the potential for repeat events.

[1] For very large earthquakes, full time-stepping simulations of the rupture process and ground motion time histories can and have been modeled. These are computationally intensive, however, and generally not state of the practice for probabilistic risk models beyond inclusion of a few selected events.

By contrast, hurricane risk is linked to a moving phenomenon that can shift over several days, so the resultant hazard dynamics can change over the life of the event. In other words, the wind speeds at location X from a specific hurricane will vary as the storm moves ashore and eventually dissipates. As a result, hurricane hazard data are organized by event as well as location.

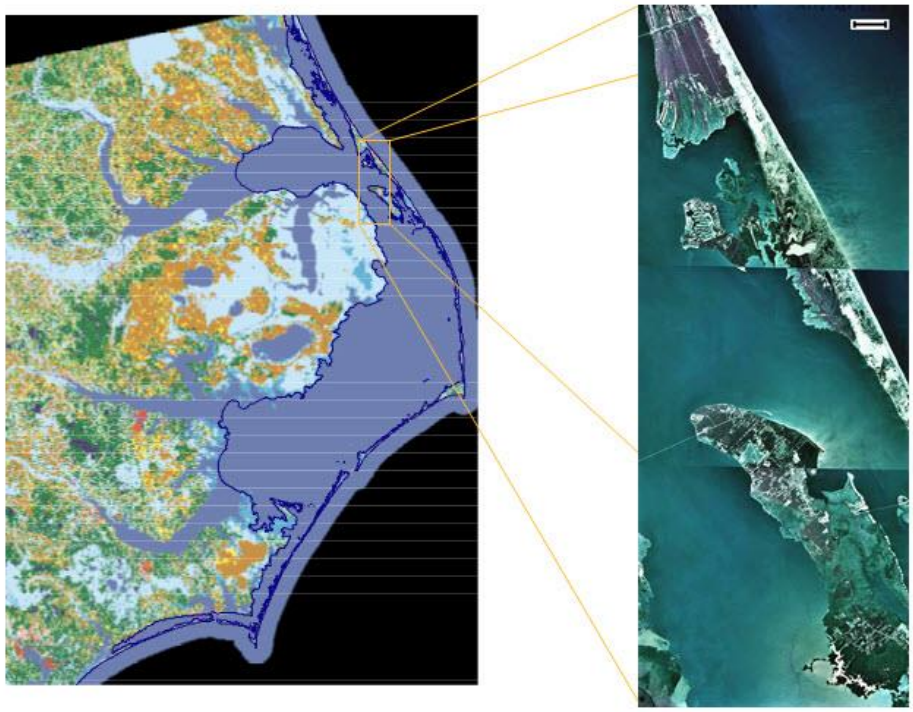
Early hurricane models were based exclusively on relatively simple parametric models with localized adjustments for site-specific features. They were static and based on location characteristics without respect to the evolution of an event. Distance to coast was one of the primary inputs, surface roughness was determined at the location in question, and elevation was taken into account for surge purposes. They did not account for the nature of the event, that is, that a hurricane is a moving phenomenon and its hazard profile changes at each site as the hurricane travels.

As increasing amounts of high-resolution empirical data on windspeeds and damage became available while computing resources simultaneously grew in power and dropped in cost, there was a shift to more realistic, calibrated hurricane models. This newer generation of models fundamentally re-examined the physical processes that drive hurricanes. These models are dynamic, based on atmospheric processes and how the resultant winds interact with location characteristics over the life of the event. Windspeeds are repeatedly calculated for all points along a storm track at each time step in the evolution of the event. In these models, surface roughness cannot be represented by a point value at the site because the actual impact of the local terrain is a function of the direction from which the wind is coming and the materials over which it has traveled.

In Unit 1, we noted how both position and geocoding resolution of a location can influence the event selection for hurricane as well as the actual hazard values. In thinking about the latter of these, it should be apparent from the comments above that there is not a simple, direct relationship between site characteristics and the resultant hazard. The values retrieved for a location can provide insight, however, as long as one is aware of methodologies and interdependencies.

Distance to coast is no longer a primary input for calculating hurricane loss because in isolation it is not a key determinant in estimating wind hazard and resultant damage. In spite of this, however, it is often used in underwriting, particularly with respect to storm surge. Note in particular the coastal marshes, illustrated as light-blue areas in Figure 33. These maps show the North Carolina coast in the eastern U.S. Along most high-risk coastlines there is more of a transition between water and land than a true boundary. Distance to coast is an easy but overly-simplified metric for characterizing hazard for wind. As you can imagine, two properties with the same distance to coast can have different hazard levels dependent on the frequency of hurricanes and the land cover between the location and coast. This is true along the entire eastern seaboard.

FIGURE 33: Redefining the Coastline as a Land-Water Interface



The new methodology that came with newer generation models called for a need to use parameters other than distance to coast to better capture the dynamic nature of hurricane loss. A hurricane can come from numerous directions. The passage of a hurricane causes wind directions to change for a location. Surface roughness describes the intervening land cover that can interrupt and reduce the wind from a storm. Surface roughness can impact wind speeds from any direction.

Figure 34 shows a pinwheel used to dissect the wind modeled for a hurricane at any location (ZIP Code or VRG centroid) into eight possible incoming directions. Surface roughness is calculated for each slice outward from the center to as far as 80 km (50 miles) away, reflecting the path over which the wind has to travel. Roughness from a slice that extends mostly over water is low while roughness that extends mostly over land is high. Surface roughness is one variable that influences a wind field. Each event or hurricane has its own wind field, also related to the evolution of its central pressure, forward speed, and radius to maximum winds.

FIGURE 34: Surface Roughness and Wind Direction over Southern Florida

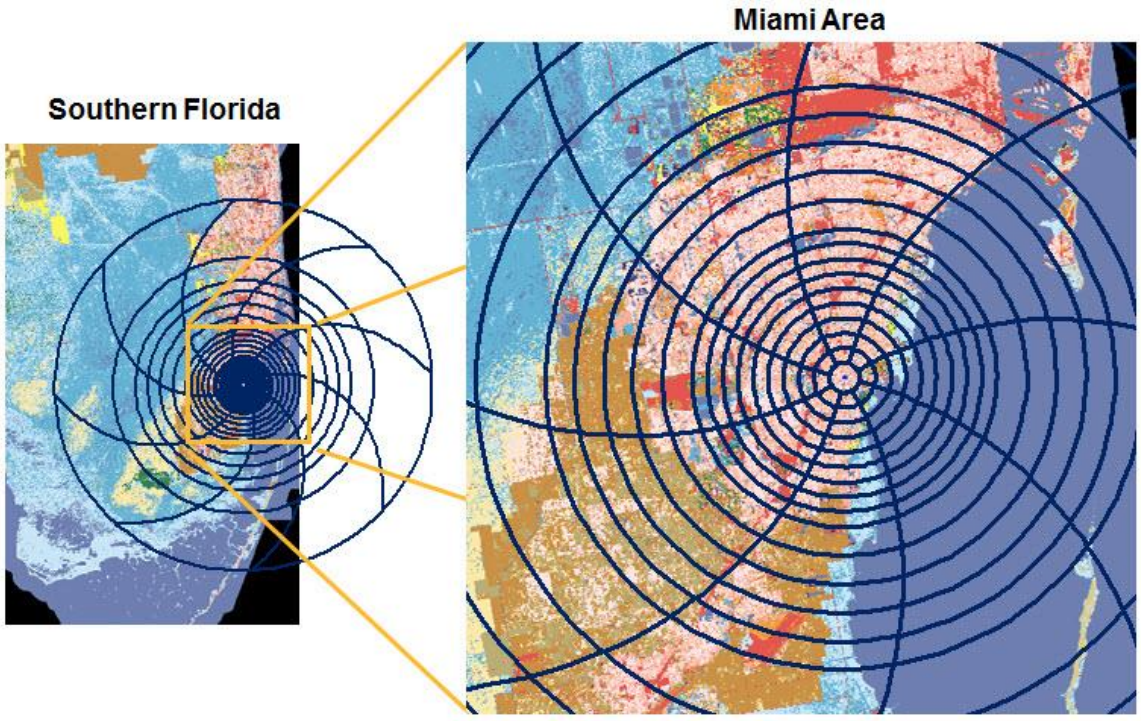


FIGURE 35: Time-Stepping Wind Field for Hurricane Floyd (1999)

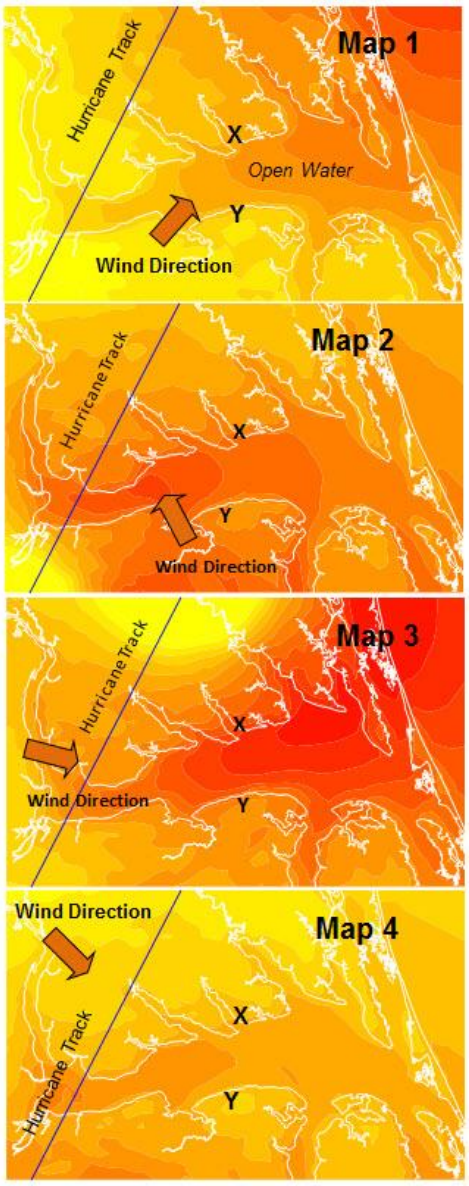


Figure 35 shows Albermarle Sound in North Carolina and the wind field for Hurricane Floyd (1999). The sequence of images shows the passing of the hurricane. The colors represent peak wind speed. Higher wind speeds are shown in darker oranges and reds. The hurricane is coming from the bottom left in map 1 and then tracks north-northeast.

Note that the characteristics of the wind field contours follow the shape of the shoreline into the sound. The points X and Y are the same distance to coast, but note the differences in wind speeds as you move from map 1 to map 4. The peak wind speed for the site is the maximum among all the time steps during the hurricane's passage.

In map 1, the eye of the hurricane is down off the southwestern corner and winds are coming up from the southwest. Point X is exposed to higher wind speeds than Y because the wind affecting Y is being slowed down by greater surface roughness. The area just south of X is all open water.

In all cases, the wind speeds affecting Y will be somewhat lower than at X not because of distance to coast (which is the same) but because of wind direction and surface roughness.

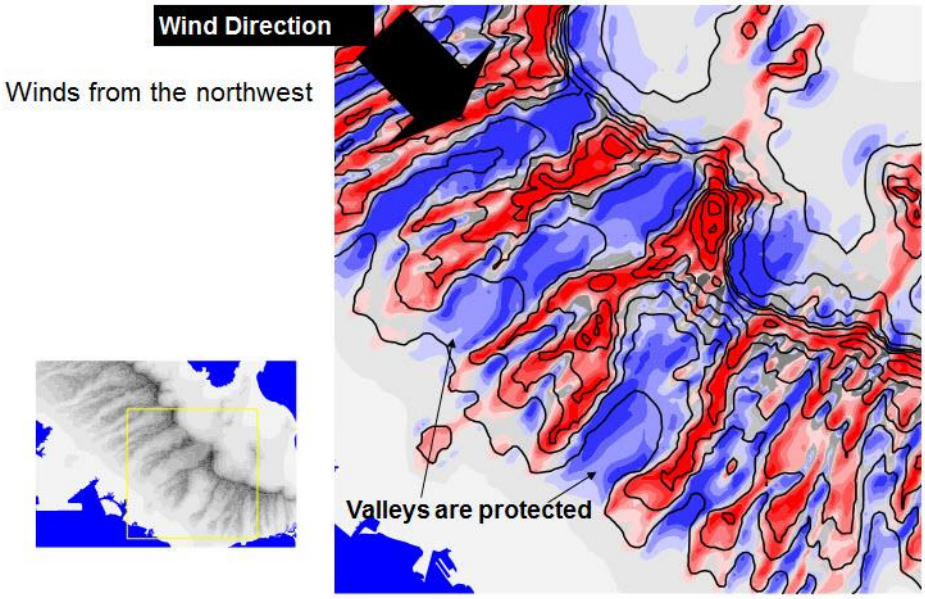
Hurricane Modeling: Topographic Effects and Wind Direction

In Florida there is little variation in topography so surface roughness remains the most influential variable impacting wind speed.

The islands of Hawaii and the Caribbean are mountainous, and wind tends to accelerate as it crests steep hills. These locations have a much greater variability in wind speed than locations with flatter topography (e.g. the gulf coast of Florida).

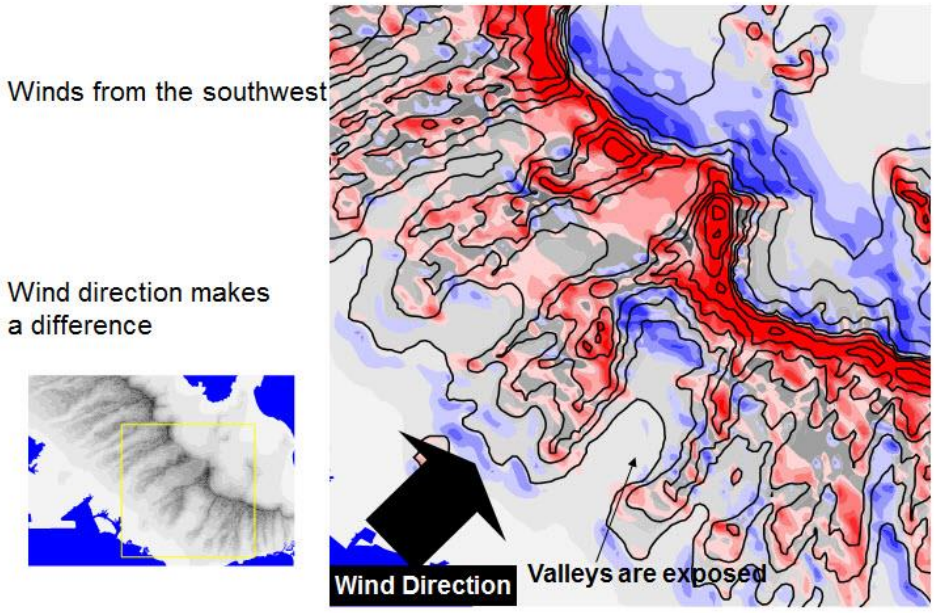
Figure 36 shows the southwestern coast of the island of Oahu in Hawaii. Lower wind speeds are in white and blues, higher wind speeds are in reds. If the wind comes from the northwest the valleys are relatively protected. The highest wind speeds are mostly confined to the ridges.

Figure 36: Wind from the Northwest, Oahu, Hawaii



If the wind comes from the southwest the winds are very high over the mountain crest so only the leeward side is protected, as shown in Figure 37.

Figure 37: Wind from the Southwest, Oahu, Hawaii



As we saw with distance to coast, elevation and the surrounding topography are relevant characteristics of a site's exposure to wind hazard, but the dynamic nature of hurricanes ensures they are not sufficient on their own to estimate the risk.

Unit 4: Closing Key Concepts

Typically, detailed hazard data are not available over large areas. In some cases, nationwide coverage is available; however, the highest detail data is typically available only for areas with the highest exposure. For example, there is nationwide soil coverage available in the U.S.; however, the highest detail data are only available for the most earthquake exposed areas, such as California, the Pacific Northwest, and larger cities in the New Madrid region.

Hazard data resolution can never be higher than geocoding resolution. So as increasing amounts of detailed hazard information are gathered, we will see an increase in the benefit of obtaining detailed geocoding resolutions.

Site hazards are determined by the physical characteristics of the particular location. Event hazards are determined based on the interaction of the event with the location over the entire life of the event.

Exposure weighting provides an approach to aggregating hazard information that improves the representation of these data at lower resolutions. This consists of a combination of unweighted hazard data (e.g. soil or wind speeds) combined with land cover data, which are generated using satellite imagery. RMS combines the land use classes with the hazard values in a GIS to calculate exposure weighted values for each ZIP Code or other aggregate unit. While this method improves the overall modeling accuracy, the overall uncertainty remains higher than when a high-resolution value can be retrieved.

Unit 5: Impact of Hazard Data on Loss Results

In this final unit, we will examine the influence of geography on hazards and the variability of risk by resolution, walk through an example of geographic risk differentiation, and look at geocoding business applications.

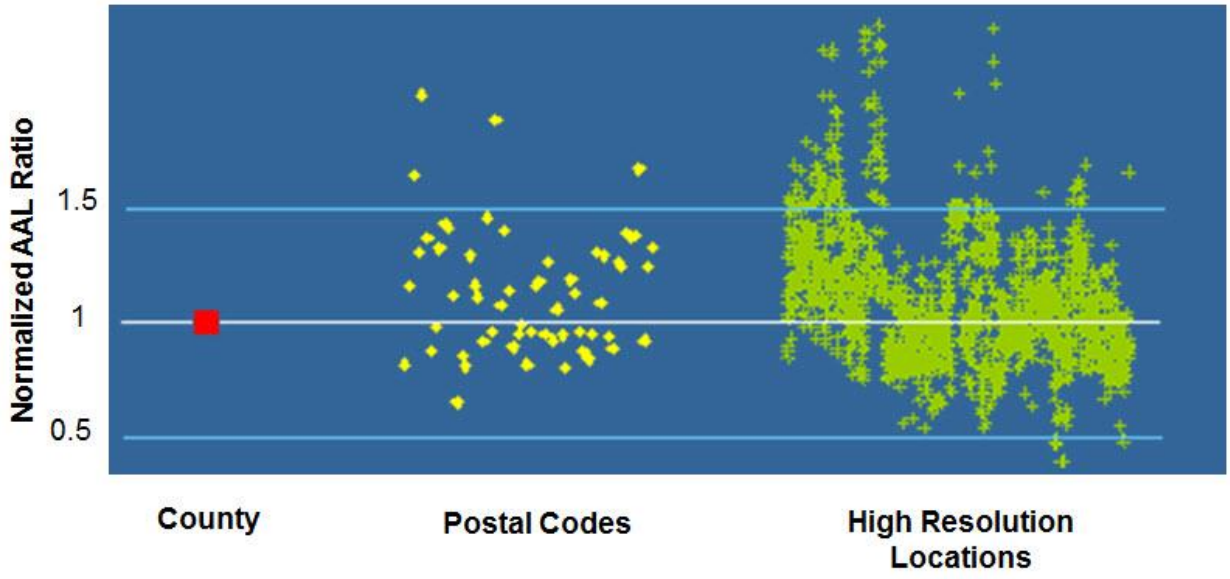
Learning Objectives:

- Understand the role of event hazards in loss outcomes where site hazard information is inconclusive or unavailable.
- Understand how spatial resolution can be used to differentiate risk.
- Apply geocoding principles to specific business problems.
- Describe some ways to integrate geographic principles into catastrophe modeling.

Variability of Risk by Resolution

The finer the resolution of geocoded locations, the greater the potential for variability within a single geographic unit. Figure 38 illustrates this concept. The graph shows the normalized average annual loss ratio (AAL ratio) of losses for three different analyses of the exposure within a single county using the same catastrophe model. Note that the normalized AAL ratio is the ground-up AAL for each location at the various geocoded levels divided by the county level average annual loss. The chart shows that the degree of variability in the results increases at higher geocoding resolutions.

FIGURE 38: Variability of Risk by Geocoding Resolution



If you have 1,000 locations that are identical in terms of their characteristics (e.g. construction, occupancy, etc.), and they all geocode to the county level, all of them will fall on the red box and have the same normalized AAL ratio. In other words, there will be no variability due to distance

from events, site effects, and such. If those same 1,000 locations can be geocoded to the postal code level, you will see some variability as shown by the yellow diamonds.

If all 1,000 locations can be geocoded to unique street addresses, you can have significant variability in the individual results. This is actually a good thing – it demonstrates that your losses are as reflective of reality as possible and it differentiates the loss potential of each risk.

Geographic Risk Differentiation Example

We know that risk can vary by geography or geocode resolution, so quality location data are critical. Next we will look at a hypothetical scenario. A multi-national corporate client employs a staff of 13 to be directly responsible for improving their data quality. This dedicated-staff works directly with producers to acquire more complete data, clean existing data, and check address validity. They ask the question: Where do we focus our efforts so we can work more efficiently?

The general response is to focus on the areas of greatest risk. The company will gain the most benefit by focusing on the areas with the highest concentrations of exposure, the greatest variability of hazard, and the highest resolution of underlying hazard data.

In this example, it would be beneficial to perform some research to determine the company's area of sensitive exposure, focus your resources on the areas that will have the greatest impact and the highest return, and develop business practices that include obtaining good data quality before your company takes the risk onto the books. For example, underwriting guidelines could be placed to require all new accounts in the enhanced data collection zones (i.e. the zones that will have the greatest impact and highest return), to include correct and validated street level address information.

Units 1-5: Closing Key Concepts

Let's review the key themes we have explored throughout these five units. None of these concepts are unique to RMS models, and they are important no matter what modeling technology your company uses. If your company licenses multiple models, you can use this information to help you with your understanding of all loss results. If you find discrepancies between products, mine your data to determine if and how geocoding and hazard retrieval have influenced your results.

Geocoding Key Concepts:

- Geocoding impact on loss results
  - Detailed Modeling
  - Aggregate Modeling
- Geocoding accuracy
- Consequences for data quality
- Geocoding and peril resolution
- Geocoding technology
- Making sense of unfamiliar address systems

Hazard Retrieval Key Concepts:

- Data resolution versus coverage
- Model changes over time
- Geography of risk
  - Earthquake
  - Windstorm
- Site hazards versus event hazards
- Business applications

Now that you have completed the reading for the Geocoding and Hazard Retrieval course, please move on to the please move on to the Geocoding and Hazard Retrieval Exercise in the CCRA portal of Owl.

Once you have completed the exercise, you must take the online assessment for this course, Geocoding and Hazard Retrieval Self-Assessment, which is found in the CCRA portal of Owl. This course will show a status of "complete" after you have viewed all the course materials and scored at least 60% on the self-assessment. You will not be able to move on to the next course until all components of the Geocoding and Hazard Retrieval course have been completed.