CCRA® Training Program

Exposure Data Analysis

September 5, 2019

Exposure Data Analysis

This CCRA course provides detailed information on exposure data quality relevant to catastrophe modeling. The quality of exposure data impacts all components of catastrophe modeling, from geocoding to modeled loss estimate, and therefore the business decisions made based on those results as well. The course will include the following units:

- **Unit 1: Introduction to Exposure Data** on page 3 – Focuses on defining and developing exposure data types of exposure data analysis and understanding the most important rules of exposure data analysis.
- **Unit 2: Exposure Data Management** on page 14 – Describes the application of exposure data analysis in catastrophe risk assessment, the downstream effects of data assumptions, and the assessment of exposure data implications on modeled loss estimates.
- **Unit 3: Assessing and Measuring Exposure Data Quality** on page 22 – Identifies the most important data quality issues, defines data quality categories, and explores ways to identify data quality issues.

Unit 1: Introduction to Exposure Data

This unit will focus on the basics necessary to better understand exposure data and the importance of data quality.

Learning Objectives:

- Define and understand the process of developing exposure data
- Understand the two rules of exposure data analysis
- Learn how to avoid common pitfalls when working with exposure data
- Identify the key concepts for each type of exposure data
- Understand exposure best practices

What Are Exposure Data?

Exposure data are all the information elements used to describe what is re/insured, how it is re/insured, where it is located, and how much it is worth. These are all critical pieces of information required to assess one's exposure to and loss potential from a catastrophic event. In addition, the exposure data also include site hazard data such as topography, soil type and elevation, and geocoding results.

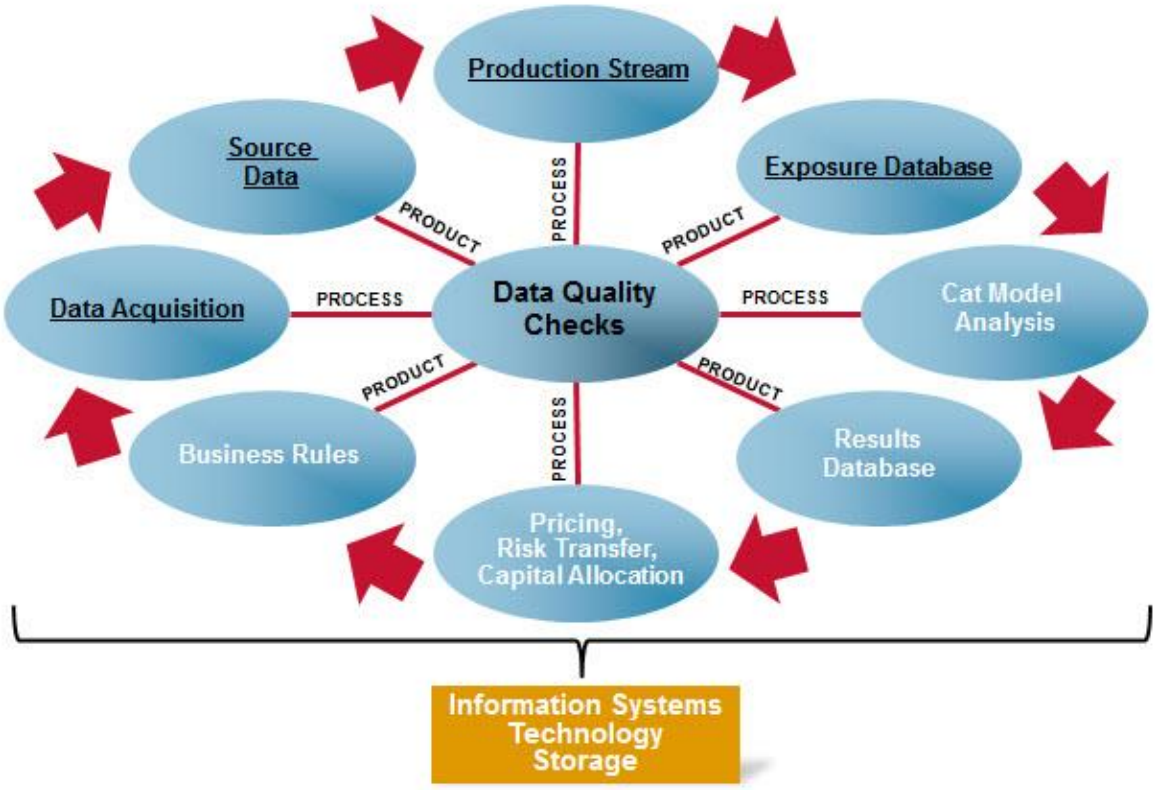
Exposure data can be stored in many formats such as spreadsheets, digital flat files, and databases. For users of RiskLink, the exposure data is contained in the exposure data module (EDM).

Understanding your exposure data is the first step to catastrophe risk management.

Lifecycle of Exposure Data

Catastrophe exposed data can be derived from several data products and processes. The comprehensive management and analysis of exposure data include cleansing, analysis, and organization. This course addresses only the processes and products in black text in Figure 1, which are the processes and products used in the creation of the EDM. It includes the impact of systems processes on exposure data throughout its lifecycle. In addition, it includes data quality checks at all points during the lifecycle.

FIGURE 1: Exposure Data Lifecycle



It is important to note that each of these processes impacts exposure data quality, an assessment of which is important in the creation of information. In addition, information can be extracted at any point in the acquisition, production, and analysis of exposure data. In a perfect world, the quality assurance process occurs at any point that information is entered, so that only quality data is in the system, and at any point information is extracted.

Everyone that touches the data throughout the cycle should understand the ultimate end goal for that data so they can make appropriate decisions where necessary throughout the process. It is important to make sure the data acquired are not only of sufficient quality for the type of analysis that will be done, but also that they are the best quality possible given business constraints. For example, if you are planning to run a hurricane analysis on a Florida residential book, then the need for data on square footage is critical to communicate not only to the owners or producers of the exposure so that this information is captured to begin with, but also to the database administrators so that the data are accessible. It is the degree of communication between the systems/IT staff and business analysts, as well as the existence of agreed upon structures and standards that can make the difference in the quality of your exposure data analysis results.

Exposure Data Analysis – The Rules

Exposure data analysis is the process of systematically applying statistical and logical techniques to describe, summarize, and compare data. The main exposure data analyses with respect to catastrophe risk management are data interpretation and data profiling, which we will discuss in more detail throughout this unit.

There are two rules of exposure data analysis:

Rule 1: Understand your end goal and end user.

Rule 2: Analyze data with a "fitness for purpose" framework.

Rule #1: Understand your end goal and end user.

In order to apply this rule, it is helpful to consider a few key questions:

- What do you want or need to do with the exposure information or the modeled results?
- How can you most efficiently analyze the information to achieve your goal?
- Who will be using these data once they are analyzed?

Take time to reflect on these questions *before* you invest significant time and effort setting up your data. Ignoring this rule may result in re-analysis, which can cost a significant amount of time and effort. Focusing on the wrong goals reduces your competitive advantage.

For example, ABC Insurance Company writes large commercial policies with policy-level limits that cover locations across Florida in the United States (i.e. blanket limits). The company wishes to understand its aggregates by postal code in the coastal counties to make sure they have not exceeded their capacity. They have asked you to compile a report to help them with the analysis. In order to create this report you will need answers to the following questions:

- What are the company underwriting guidelines? Is there a maximum limit per postal code in the coastal counties? If there are postal codes over this limit, it could mean that the limits are not being captured properly or that underwriters are not adhering to the underwriting guidelines.
- What is the resolution of the data? If the data were captured and/or geocoded to a coarser resolution than postal code (e.g. county), then not all the data will be able to be captured and analyzed at the necessary level for reporting, and the resulting analysis may be inaccurate.
- Are the data complete? The data must include all policies in force in order to get an accurate assessment of the postal code-level aggregates.

- What parameters need to be aggregated? Is the company just looking for total insured values by postal code, or is it total insured values by line of business and by construction type as well?
- What financial structures are in place? The fact that you have blanket limits that may cover locations in multiple counties means that you will have to determine how much limit exists in any given postal code for monitoring aggregates. Also, is there any facultative reinsurance in place that would reduce the company's exposure?

Rule #2: Analyze data within a "fitness for purpose" framework.

The second rule means that the analysis type should be appropriate to achieve the goals and objectives set out by the end user.

Exposure data analysis requirements are not uniform between:

- Perils
- Regions
- Market Segments (Line of Business)

Peril models use different data resolutions depending upon the resolution of the model, the level of hazard in the region, and the market segment or line of business being analyzed. For example, a peril model with a high hazard gradient (i.e. the hazard varies greatly within small areas), such as river flood, requires high resolution data to most accurately determine the exposure to this peril. By contrast, European windstorms generally have low damage over wide areas, and therefore, a coarser resolution of geocoding would be perfectly acceptable for analyzing the exposure to this peril.

Another example is the western region of the United States. If running an earthquake analysis in California, it is important to note that the soil types change rapidly over short distances in the western half of the state, but in the eastern half, the soils are more uniform. With this in mind, having only postal code level data available for locations in the eastern half of California may be acceptable for certain types of analyses. In the western half of the state, however, detailed location information becomes very important in order correctly assess the site hazards.

In addition, the type of information being analyzed will be different for residential vs. commercial lines of business. A single family home located in a flood-exposed area will be at risk of flooding regardless of the number of stories. However, a commercial risk in the same zone may not be exposed to flooding at all if the insured is located above the first floor of the building. In this example, having accurate information on the number of stories and the floors occupied is important information to accurately assess flood exposure.

Exposure Data Analysis Best Practices

In addition to having a clear understanding of the end goal of your analysis, you should also have a clear understanding of the necessary data and analyses required for the stated business purpose. Understanding the business value of the knowledge you are generating will also provide the framework within which you analyze the data.

When formulating insights based on the interpretation of data and information, it is best to first identify what is known, i.e. the “facts.” Once this is established, you can identify what important data you do not have or do not understand in the context of your analysis goal, and then request additional information. If it is not possible to get this additional information, then some assumptions may have to be made and documented in order to fill in for missing data. Exposure data assumptions will be discussed later in this unit as it should be approached with its own set of best practices. At the very least, assumptions should be applied consistently and knowledgably. Finally, there is the critical step of documentation -- a best practice that should be applied in every analysis and yet the easiest step to dismiss or put off when deadlines are looming.

Common Traps in Exposure Data Analysis

Knowing the common exposure data analysis traps and understanding how to avoid them will not only improve your analysis, but will also serve to reinforce the rules discussed in the previous sections.

One of the most common traps already mentioned is a lack of documentation, which can often lead to misinterpretation of the data being analyzed. A lack of documentation could include codes for construction types that are not defined, or failing to record the process used to fill in missing coverage limits in the data. A related trap is not having a complete understanding of the methodology being used in an analysis, such as understanding the spider accumulation methodology. In addition, determining which data hold the greatest business significance and will offer the greatest competitive advantage can also be a challenge. For example, a company may decide to cut costs by excluding some data from the data acquisition process. However, if the company does not understand which information holds the greatest value, an opportunity may be missed and important data may be excluded that would otherwise support a current or future exposure data analysis.

Another trap is not knowing the right questions to ask regarding the data with which you are working. For example, being unfamiliar with a particular peril could result in data that is not at the required resolution for that peril. In addition, making the wrong assumptions regarding exposure data can lead to an inaccurate analysis. Prioritizing data issues correctly can also be a problem, as can a lack of foresight and flexibility. While some of these examples are traps that can be avoided, common data problems still exist and must be dealt with accordingly.

The common traps listed here are not a comprehensive list, but highlight some of the issues and decisions faced by catastrophe risk analysts. This course will address some of these traps

throughout the remainder of the unit as we shift the focus to the types of exposure data analysis.

Types of Exposure Data Analysis

In this section, we will discuss two types of exposure data analysis: **data interpretation** and **data profiling**.

The interpretation of exposure and results data occurs at several points in the analysis process. It includes everything from how to code a construction of “wood frame with masonry” as listed in an underwriting submission, to interpreting whether a quoted risk that marginally meets underwriting guidelines is worth booking.

Exposure data profiling is the grouping of exposure data by one or more parameters. Using Rule 1 of data analysis, how you organize your data depends on the end goal and end user. Examples of exposure data profiling include grouping or aggregating data; creating reports, graphs and charts; and the creation of maps.

Data Interpretation

Data interpretation can sometimes be an automated process and include data content and format validation. An example is the automated interpretation of parameters in underwriting submissions and schedules of locations for input into a “master” exposure database, such as the RiskLink EDM. Automated interpretations require the application of knowledge-based information in a structured manner; however they can also be inflexible and can become outdated.

In contrast, data interpretation can be conducted by individuals. This requires specific knowledge applied on a case-by-case basis. One example is deciding what, if any, assumptions should be applied to unresolved or incomplete data.

A third variant of data interpretation is the assessment of data based on heuristics and/or data comparisons. In other words, this is the interpretation of data based on observed patterns, simplification, or educated experimentation that serves to fill in information or processes where details are lacking.

Lastly, the interpretation of exposure data and results data analyses is not always a process that can be structured to apply uniformly across all lines of business. Different business issues require different exposure data interpretation concepts and applications.

Example of Data Interpretation – Coding Complex Policy Structures

Anyone who has received a submission slip or other information regarding a policy has most likely experienced some uncertainty regarding how to code it for use with RiskLink or another software application. Oftentimes the language and structure of the policy do not fit completely within the definitions and structures defined by the database format or the selection options for the software. It is important to resolve this mismatch in such a way that you are confident that

the exposure data as represented in the EDM is also representative of the exposure from the perspective of the insurance stakeholders.

If you are the person interpreting or working with a populated exposure database, be aware that interpretations were made when coding information. If these are not documented, then they are not easily tied back to the submission and the judgment used in the decision-making process is lost.

When working with complex policy structures, it is important to note exceptions mentioned in the submission, such as windstorm attachment points that are for named storms only. Deductibles and limits must be interpreted correctly for each peril. For example, a tornado deductible may apply separately from a hail deductible. The analyst must decide how to best represent this in the exposure data processed by the model. Awareness of the process used to input data is critical to the interpretation of downstream analysis results and the business decisions ultimately made.

Example of Data Interpretation – Aggregate Exposure Data

Another example is the interpretation of aggregate data. Aggregate exposure data is a group of exposures categorized as one record. Aggregate data can be represented as limits by a geographic region, as aggregated locations within policies, or as values grouped by primary data characteristics. There are multiple ways to estimate loss from aggregate data, including the use of market share analyses, aggregate analysis tools such as RiskLink ALM, and disaggregation of the data for use in a detailed loss model.

When working with aggregate data, it is critical to identify what is known about the data. Important information may be populated in the database, such as the “number of buildings” or “count” field for each aggregate record. This is critical information as $10M of total insured value (TIV) from only one property is not equivalent to $10M of TIV from 100 properties in terms of loss potential.

When multiple buildings are aggregated or grouped, it may require that some assumptions be made about the average primary data characteristics for all the buildings. There are several approaches used to assign a single value to key characteristics including using average values, “conservative” values, or the characteristics of the “dominant” or primary location. This primary location can be chosen based on its majority contribution to the TIV, amount of information known about the location, its size, or its association with the listed address.

One of the key characteristics that is lost when data are aggregated is detailed, site-specific geographic/address information. Best practices emphasize understanding not only the business value, but also the fitness for purpose of your exposure data. In this context, it is valuable to assess the geographic resolution data in light of the model that you are using.

Finally, detailed financial information is lost, including coverage information (building, contents, and time element), deductibles, and/or limits. For example, a $25,000 deductible on a single $10M property is not equivalent to one-hundred $100,000 properties in a common geographic space that each have a $250 deductible.

Unfortunately, aggregate data sometimes represent the only level of information available. While not ideal, there are some approaches to working with this information, as will be discussed throughout this course. Regardless, when receiving aggregate data, the preferred option is to obtain, if possible, the detailed data from which it was created.

Data Interpretation – Unknown Data

When values are unknown or missing in exposure data, one approach for replacing the unknown data with valid assumptions is to compare them to similar data. Inventory databases may provide the necessary information. These databases have been analyzed to identify trends and averaged data values in the industry. If the business is similar to the industry, then utilizing an inventory database provides a reasonable alternative. As an example, if construction class is unknown, then a construction inventory database is available in RMS catastrophe models that will replace the missing information in an analysis with weighted averages of construction type appropriate for the occupancy and year built in a that area. However, if the exposure is a niche (i.e. specialty) market, then some alternative assumptions may be more appropriate. For some perils or regions, such as terrorism, there may be no accurate inventory.

Other databases may be available for a region to provide missing information or as validation for existing data characteristics. These include building-specific construction information (e.g. Sanborn data or RMS ExposureSource) or company claims databases.

Unknown information should be minimized no matter the peril or market sector since the true characterization of risk is quickly lost if this default becomes commonplace. Although some degree of unknown information will always exist, the key is to continually apply pressure to the sources of data acquisition to determine the actual, accurate building characteristics. Finally, it is important to understand a company's best practices for handling incomplete exposure data.

Data Interpretation – Data Assumptions

No exposure database is perfect; therefore, it may be necessary to make assumptions about data that are unknown or seemingly incorrect. Assumptions can be based off of other industry databases, peril/region/line of business averages, or via data trends and known data relationships. A simple example of the latter is the reasonable assumption that most northeast Florida residential buildings are wood frame.

A critical practice when making data assumptions is documentation (for later referral), justification (to explain reasoning), and verification (to check for validity).

While data assumptions are critical to enhancing and improving incomplete or poorly resolved data, it is important to remember that when data interpretation and assumptions are made, they are made *from* data and are not data themselves, and therefore they must be clearly documented as such. Following are some data assumption examples that *improve* data quality:

- "Based on building codes, I know that after 1988, all equipment in CA is well braced. Therefore, if my year built is 1990, I will select the 'generally well braced' secondary modifier as the default."

- "I received this data from a specialty commercial insurance company, but there is no occupancy data. I will select 'General Commercial' as the default."

Data assumptions certainly have the potential to degrade the quality of data if poor extrapolations are made, or if assumptions are made from outdated or incorrect information. Examples of assumptions that *degrade* data quality include:

- "Based on common construction practices, for all locations built after 1976, many companies are entering 'Properly Anchored' for the roof/wall connections." - While many roofs/walls are properly anchored, not all are, therefore it is better to leave this characteristic as unknown.
- "I've always defaulted all houses to 1 story. When I started this job, I was told that for residential buildings 1-3 stories all perform the same in the catastrophe model." - There was a time when RMS models treated all houses of 1-3 stories the same. Now, however, they are treated differently. The process of generalizing to save time, although once valid, will now yield a potentially misleading result. This is one reason why it is important to stay up to date on model changes and assumptions being used.

Some of the assumptions may actually bias the modeled loss results to a more favorable outcome (depending upon your perspective); however, a "favorable" outcome does not equate to improved data. As an example, adding a secondary modifier for earthquake risk that assumes equipment is "well braced" will decrease losses slightly. Making this type of assumption may require further verification that this is valid.

Data Profiling

The next type of exposure data analysis we will review is data profiling, or the analysis of exposure data categorized by one or more parameters.

One type of exposure data profile is the grouping and aggregation of data, which results in a loss of data detail. This may be done to arrive at the sum of data values, or to improve analysis performance by reducing the number of "locations" modeled. It may also be used to simplify the characterization of complex or poorly characterized risks, such as a campus, or to reduce the size of the exposure database.

Reporting, graphing, and mapping are visual representations of exposure data profiles. The type of profile you use and how you organize the exposure data will depend on your end goal and end user and your fitness for purpose.

There are many challenges with profiling such as how to handle multi-location policies with blanket financial terms. Policy blanket terms must be accounted for when applied to locations or coverages that are not all covered by the same parameter against which the data is being profiled. The challenges of determining regional limit contributions, whether for cost allocations or accounting for regional sub-limits, are especially daunting. Another challenge is accounting for attachment points and deductibles. Assume you are allocating regional limits for an account with the following two policies:

- Policy #1: $5M limit that attaches above $1M
- Policy #2: $5M limit that attaches above $100M

Your regional profiles would produce exactly the same values for both even though it is clear that the loss potential from a given event is very different for each policy. In this case, it is very important to consider attachment points in some way. There are several valid options to solve the issue; however, all options will require that some assumptions be made and therefore, documenting the process is critical to understanding the output.

Finally, it is often falsely assumed that analyses that are based on aggregate data yield more conservative (i.e. higher) loss results than detailed data analyses. In fact, it is impossible to predict whether aggregate data are conservative in nature without supporting documentation. An example is an assessment of multi-peril risks, where one construction type may provide worst-case estimates of potential loss for one peril, but not for another. One additional pitfall to be aware of is using small populations of data to create average values for aggregate data. The end result is that your averages will be biased.

The following e-learning interaction highlights an example of creating a data profile using three different methods.

E-LEARNING INTERACTION 1: Data Profiling

To view an interactive example of mapping exposure data by county using data profiles, go to the Exposure Data Analysis course in the CCRA portal of Owl and select Interaction 1: Data Profiling.

Imposing the #1 rule of exposure data analysis, “understand your end goal *and end user,*” is critical before starting the exercise highlighted in the previous e-learning interaction. Is the map a regulatory reporting requirement, or is it being used for internal capital allocation purposes? An up-front understanding of what the gross exposure by county map is being used for may save you much time in the end as one approach may be more appropriate than another for various reasons.

Unit 1: Closing Key Concepts

Understanding the end goal and end user is an important aspect of any exposure data analysis. In addition, keeping in mind the fitness for purpose of the data and the analyses for the peril, region, and market segment being analyzed are equally important.

Understanding the end goal of analyzing data is critical before starting to perform exposure data analyses. Most analysts have experience re-running analyses because the correct information was not output, or the correct data were not used. Questions that you should be able to answer before beginning data preparation for analysis or modeling are:

- What do you want or need from the data once it is modeled?
- Do you know what other downstream uses of the modeled data are intended?
- How can you most efficiently get the right output to achieve your goal?
- Do you understand what the data parameters represent and their definitions?
- Do you have a clear understanding of the business decisions that will be made based on the analysis results?

Always take time to reflect on these topics before you invest significant amounts of time and effort in setting up your data.

Finally, exposure data profiles should not be used as predictors of actual loss potential. There are many challenges associated with interpreting exposure data and creating exposure data profiles that are reflective of loss potential. If correctly documented, justified, and verified, data assumptions can be valuable for creating more accurate and informative exposure data profiles.

Unit 2: Exposure Data Management

This unit will focus on the basics necessary to better understand exposure data management and how exposure data affects other aspects of modeling.

Learning Objectives:

- Describe the applications of exposure data analysis in catastrophe risk assessment.
- Understand the downstream effects of data coding and data assumption choices.
- Assess exposure data implications on modeled loss estimates.

Exposure Data Applications

We will focus on the following eight common exposure data business applications that are used in the insurance industry. Please note that this is not a comprehensive list.

- Catastrophe model data preparation
- Conforming to data and process standards
- Assess exposure to non-modeled perils
- Pre- and post-event response
- Real time communication
- Risk evaluation and comparison over time

This unit details each of these applications. Specifically, we will provide a description of the types of analyses used for each of the applications and a discussion of examples and/or potential issues.

In addition to the six previously listed, exposure data quality measures and improvement are covered in Unit 3, and accumulation management is treated in a separate course. These applications have been subject to an increased industry focus since the September 11, 2001 terrorist attacks and the 2004 and 2005 U.S. hurricane seasons. The resolution, completeness, and accuracy of exposure data were significant factors in the difference between actual and modeled losses for these events. In addition, non-modeled losses contributed significantly to overall industry losses, making accumulation management, or the understanding of exposure concentrations, a critical aspect of catastrophe risk analysis.

Catastrophe Model Data Preparation

Before beginning any type of exposure data analysis, it is important to understand what data you have as the foundation for understanding what you need and properly interpreting results. Knowing the reporting and output needs prior to running a catastrophe model is another necessary step to make sure the proposed goals of the analysis are met.

Questions that you should be able to answer before beginning data preparation for analysis or modeling include:

- What do you want or need from the data once it is modeled?

- Do you know what other downstream uses of the modeled data will be?
- How can you most efficiently get the right output to achieve your goal?
- Do you understand what the data parameters represent and their definitions?
- Do you have a clear understanding of the business decisions that will be made based on the analysis results?

Always take time to reflect on these topics BEFORE you invest significant amounts of time and effort in setting up your data.

Let's look at an example where data are prepared for a catastrophe model analysis. A good starting point is to understand what lines of business the exposure data covers. This will provide critical information on what type of building characteristics need to be captured, as well as what type of construction schema should be used. The perils that are covered also drive the type and resolution of information required.

If you determine that more information is needed prior to modeling, it is important to know where to find the data and whether there is any documentation on data assumptions that were made. Finally, it is not always appropriate to assume that the location information represents the intention of the policy wording. Are the values really limits? Is the county-level geocoded information based on aggregate or multi-location data? Are the deductibles and limits coded correctly and at the right level? Many of these questions overlap with data quality concerns to be covered in subsequent course material.

If you are conducting an earthquake analysis and the only modeled output needed is U.S. earthquake average annual loss (AAL) results by county, then it may not be worth your time to run any earthquake analyses other than an exceedance probability (EP) analysis. If you have been asked to provide losses by location, it is necessary to turn on location level output. This can dramatically increase analysis run-time, and expectations need to be set with management regarding completion times.

In summary, when preparing your data for modeling, it is important to understand your exposure data. Failing to do so can lead to results that are inaccurate, misinterpreted, or both. Understanding your exposure data will help you to properly prepare the data in order to create the necessary modeled output. Always take the time to reflect on the end goal, as well as the fitness for purpose, BEFORE you invest significant amounts of time and effort in preparing the data for modeling.

Conforming to Data and Process Standards

Another common application is the preparation of data to conform to existing formats and standards such as the RMS EDM. Standards are a set of rules and guidelines that provide a common framework for communication and for structuring information. From a data perspective, it means a given data element has a common name, definition, and value set or format that applies across a whole organization, industry, etc. From a process perspective, this means

documentation on the best way to do the job. This can be a “best practices” document that provides a set of rules, validation steps, or recommendations, including data dictionaries or guidelines set by industry standards groups. It is important to consider where process standards fit into the data acquisition, data production, and systems processes both for internal and external reporting requirements. One standard does not fit all organizational processes.

Data standards structure data values into formats that can be interpreted by several groups and applications. In several standards, numbers are used to reference specific values. Using a standard allows groups to map values consistently between both old and existing data sources as long as they are properly defined. Finally, the ability to share data is a significant driver in the adoption of data standards.

There are often several possible values for one field. For example, a field may be populated with five different values that all signify the same information. Alternatively, a set of value options for a field may not fully span the options available to describe the field. Data formats and definitions change over time as new versions of the standard are released. When receiving data, it is always important to verify the version of the data (e.g. a RiskLink 16.0 EDM vs. a RiskLink 18.0 EDM).

Issues may arise when conforming to a standard. For example, the lack of universal adoption of a standard due to concerns over data access security, conversion, and maintenance costs. In many cases, a company’s proprietary systems and data are part of a profit center. Re-tooling the data and associated systems is viewed not only as cost-prohibitive but also as reducing its value as a profitable and marketable product.

Finally, determine if the standard is appropriate for the purpose of the data. The "same" data from more than one file or table might not be able to be merged into one table because the parameter types and formats do not match. The capture of codes may be meaningless for some users, and limits the utility of the database for the intended purpose.

Assess Exposure to Non-Modeled Perils

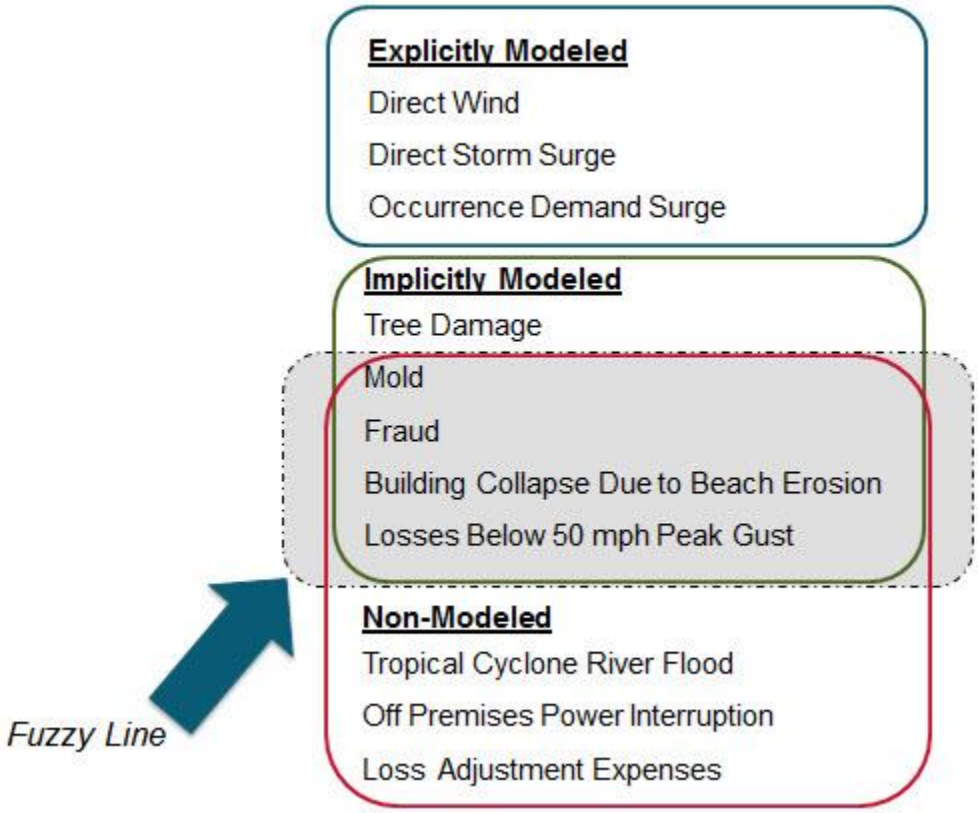
Data profiling is increasingly being used for those risks exposed to non-modeled perils in order to monitor capacity, allocate costs or capital, and assess potential losses.

Events such as Hurricane Katrina in the U.S. (2005) have dramatically highlighted the need to analyze exposure for windstorm-induced flood due to rainfall, urban drainage failure, or urban flood defense failure. Particularly high, non-modeled loss sources for Katrina included mold due to delays in accessing locations, flood loss mixing with wind settlements, Business Interruption triggers not tied to physical damage, and specialized occupancies not covered in the vulnerability model at the time of the loss (e.g. floating casinos).

Examples of modeled and non-modeled windstorm losses are listed in Figure 2. Explicitly modeled hazards are those whose damage can be tied to a specific hazard via claims information and laboratory or scientific studies. Implicitly modeled hazards are those which, as a result of inadequate loss history, claims information, or scientific studies, are not called out as a

separate cause of loss in a model but are considered to be included with the explicit hazards. Lastly, non-modeled hazards are those which can often be identified and isolated from the other peril hazards, but for which not enough information or modeling sophistication exists to model them.

FIGURE 2: Modeled and Non-Modeled Windstorm Losses



One of the most commonly used tools to assess non-modeled losses are exposure data profiles that include assessment of the geographic concentration around the source of the hazard ("ground zero") as well as the absolute and relative contributions of the key drivers of loss to total exposed values or limits. Finally, accumulation management analyses, which are highly sophisticated data profiling analyses, are increasingly being used to understand and manage non-modeled exposures.

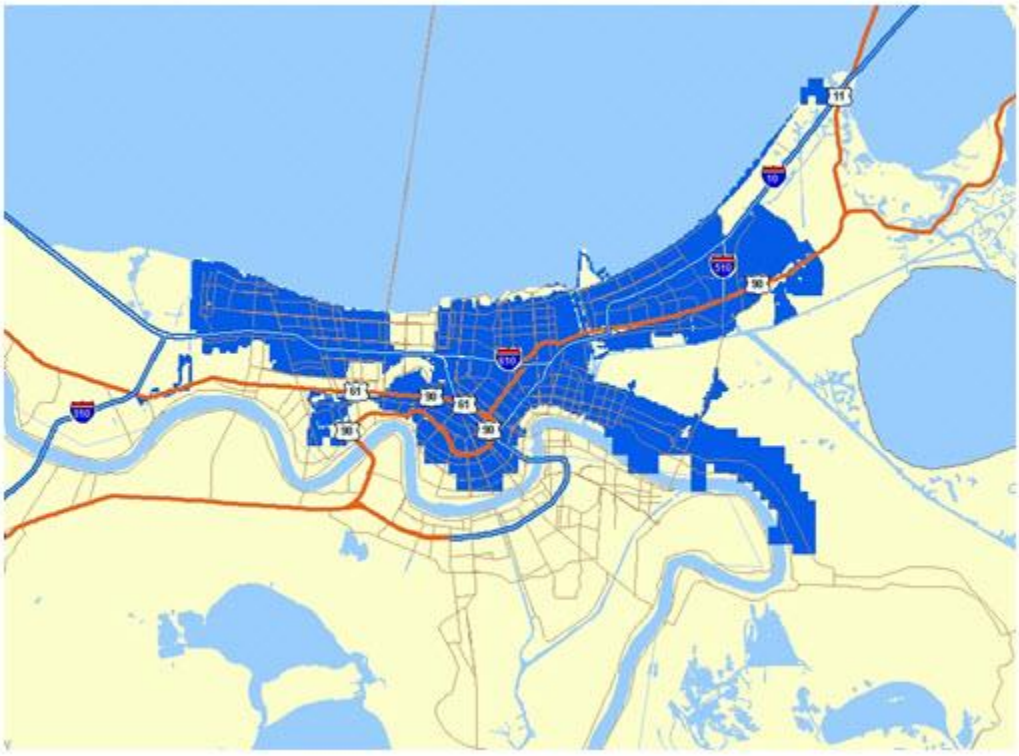
When assessing loss potential for a peril, region, and line of business, it is important to understand what is being included in the model. For example, a windstorm loss model may include storm surge in one region and not another. Different regions also require the assessment of different drivers of loss. For this reason, it is important to identify the data needed for profiling, in terms of both data availability and peril perspectives. Finally, it is important to understand your data resolution needs when profiling data for specific perils. Those perils with high hazard gradients (e.g. terrorism) require a detailed resolution of information to identify potential correlations of risk.

Pre- and Post-Event Response

Perils, such as tropical cyclone, extra-tropical cyclone, and wildfire evolve over time scales that lend themselves to both pre- and post-event exposure analysis for loss estimation. For those events that can be tracked in near real-time, such as tropical cyclones, it is possible to use pre-landfall analyses with ranges of possible stochastic events to understand which exposures are at risk and, to a degree, “predict losses.” Similarly, it is possible to overlay hazard maps of wildfire and flood onto exposure maps to assess potential losses, as well as where to focus claims efforts. Post-event mapping of claims losses with modeled losses or hazard maps are critical to understanding a host of issues from data quality to non-modeled loss contribution to potential claims abuses.

The map shown in Figure 3 is of the flooding from Hurricane Katrina in 2005. Using a map overlay of this hazard layer with portfolio or account exposures will help identify regions of potential losses in the at-risk ZIP Codes. This also highlights an example where a non-modeled peril, such as tropical cyclone flooding and defense failure, can be assessed using exposure data analysis.

FIGURE 3: Hurricane Katrina (2005) Flooding



Finally, one of the issues to consider when using exposure data analyses for pre- or post-event estimation is that the data and the profile should be at the geographic resolution appropriate to the gradient of the hazard. For example, postal code resolution profiles may be appropriate for a

large-scale European windstorm, but that resolution would not necessarily be appropriate for flood.

The accuracy of your analysis is directly tied to the quality of the data being used in the analysis. It is important to remember that any analyses outside of actual loss modeling with footprints or stochastic events are not predictors of loss, they are merely assessments of what risks are exposed to loss. That said, they do lend themselves to comparison with actual claims losses and provide valuable reference points for comparison to modeled loss. Thus a postal code level geographic profile of values can be compared with postal code aggregates of claims losses. If the limit exposed to an event is vastly larger than the modeled loss estimate, small variations in the modeling can generate large changes in the outcome.

Real Time Communication

Exposure data analyses are often quick and they lend themselves to near real-time communication of portfolio metrics on a weekly or even daily basis. Quick decisions must be made on the pricing and booking of risks, particularly in a hard market. Questions such as “How does my new risk contribute to my existing accumulation?”, “How are my exposures distributed?”, and “Can I monitor event exposure in real-time?” are being posed in an environment of immediacy and technological sophistication that allow for the dissemination of real-time knowledge to support decisions from the analyst level to the executive level. Network or internet access to this information via “dashboards” or “scorecards” is one potential solution that allows quick global access to time sensitive information.

One example of real-time communication is the RMS Event Response services. Individuals within an organization can have real-time access to an approaching hurricane, and by monitoring the track of the storm, they can evaluate potential areas at risk prior to landfall. While event info is readily available, there are no insured loss data before landfall. By uploading a portfolio of current exposures, users can assess the risks in the path of the storm and isolate potential loss concentrations for claims resource planning and deployment. From a geographical perspective it allows the user to identify what counties could be impacted and make an assessment of severity based on the loss by county. The storm tracks are updated multiple times a day so the real-time nature of the service is preserved.

While the RMS Event Response analyses may be recent, one must be careful to ensure that the underlying exposure data are up to date as well. Another potential issue is the dissemination of information to inexperienced interpreters of the information. Interpretation can be made out of context if the key components are not understood and decisions are made quickly. Finally, real-time communication may be limited by the amount of information that is available, thus limiting the decision-making.

Risk Evaluation and Comparison over Time

In contrast to real-time communication, exposure data analyses are often used over the longer term to provide comparisons and risk evaluations over time. The use of profiles can assist in

understanding whether changes in your exposure or catastrophe risk model changes are driving changes in modeled loss (or both!). They can also highlight data quality over time.

For example, consider the year-on-year assessment of a portfolio comparison of TIV (total insured value) by occupancy class. In order for the comparison to be valid, the TIV must be defined consistently each year. For example, the same coverages must be included each year. It is also important to make sure that what is meant by "TIV" is consistent. Is the TIV the total value, total limits, or something else? Because different individuals may think of TIV differently, it is important to make sure a clear and consistent definition is used each year. It is also important to be sure that the risks are classified consistently from year to year. If there was a change in the way the exposures were classified or in the way the system categorized the occupancies, then the result will not be an apples-to-apples comparison. Figure 4 shows an example of comparison values by construction class over a two year period. The classification system used was changed in from ATC construction codes in 2013 to RMS construction codes in 2014. Because a code of "2" means light metal in 2013 and masonry in 2014, comparing based on the construction codes in the data is not a valid comparison. Because of this change in coding, comparing the two years is not an apples-to-apples comparison.

FIGURE 4: Example of Value Comparison over Time

Construction Code	ATC Description	RMS Description	2013	2014	% Change
1	Wood Frame	Wood	$144,023	$148,071	2.8%
2	Light Metal	Masonry	$35,778	$49,620	38.7%
3	Unreinforced Masonry Wall	Reinforced Concrete	$30,254	$33,491	10.7%
9	Braced Steel Frame	Steel Frame	$46,184	$53,221	15.2%
17	Mobile Home	Underground Solid Storage Tank	$24,847	$22,575	-9.1%
0	Unknown	Unknown	$12,822	$15,459	20.6%
Total			$293,908	$322,437	9.7%

It is critical to set out the goals of your exposure data comparison at the beginning. Finding out that you need to track down and run analyses on data that are months old because of new benchmarking mandates can be arduous and time consuming. It is important to keep common data definitions over time for comparison purposes and to document those definitions for future analyses with links to the referenced data.

Unit 2: Closing Key Concepts

There are many exposure data analysis business applications, and each one has a subset of peril, region, and market segment (line of business) issues that must be addressed. This unit discussed each of the following in some detail:

- Catastrophe model data preparation
- Conforming to data and process standards
- Assess exposure to non-modeled perils
- Pre- and post-event response
- Real time communication
- Risk evaluation and comparison over time

Taking the time to ask some key questions, to understand what can and cannot be modeled, and to understand the data and the end goal/end user will save time in the end and will result in the most accurate and efficient analysis possible for the data.

Unit 3: Assessing and Measuring Exposure Data Quality

The first two units of the Exposure Data Analysis course focused on exposure data analysis types and business applications. However, if the quality of the data used in an exposure data analysis is poor, then the business decisions that are made based on that analysis may be sub-optimal to a company. This next unit will focus on the assessment of data quality in the context of catastrophe risk analysis. Where does data quality play a role in the catastrophe risk management data lifecycle? The short answer is everywhere. Catastrophe-exposed data are derived from several data products and processes. The comprehensive identification of exposure data quality issues can be categorized by the impact of data and database (e.g. product) quality, the processes that affect the data, and systems/storage.

Learning Objectives:

- Understand the factors that can cause modeled data to be different from actual exposures.
- Identify data quality issues in the context of catastrophe risk management data flow.
- Define exposure data quality in the context of its fitness for purpose.
- Identify the major data quality issues in the insurance/reinsurance market.
- Identify and define data quality categories and the methods used to identify them in exposure data.

Defining Exposure Data Quality

There is no single industry-sanctioned definition of exposure data quality. Consequently, the definition that will be used as the basis for this course is one based on the International Organization for Standardization (ISO), a federation of national standards. Their definition was reworded by Abate et. al. (1998[1]) as follows: "[Data is] of the required quality if it satisfies the requirements stated in a particular specification and the specification reflects the implied needs of the user". This definition provides a "fitness for purpose" definition, a critical aspect when addressing exposure data quality.

In addition, the Actuarial Standards of Practice (ASOP) 23[2], which covers data quality, provides a fitness for purpose definition in the context of data analysis: "For the purpose of data quality, data are appropriate if they are suitable for the intended purpose of an analysis and relevant to the system or process being analyzed."

Data fitness for purpose is the concept that data quality may change depending upon the requirements of the end user (e.g., underwriter needs vs. actuarial analyst needs), the context

[1] *Marcey L. Abate, Kathleen V. Diegert, Sandia National Laboratories and Heather W. Allen, Heather Allen & Associates, 1998, "A Hierarchical Approach to Improving Data Quality", Data Quality, Volume 4 Number 1, September, 1998.*

[2] *Actuarial Standard of Practice 23. Data Quality (Doc. No. 097; December 2004)*

of different market segments (e.g., workers' compensation, commercial, or industrial), and the analysis requirements.

What is Data Quality Assessment?

If asked to explain how to assess the quality of data in an EDM, the first task most analysts would state is querying the geocoding resolution of the locations. Many companies may even have guidelines that specify the percent of a portfolio's locations that must have street or high level geocoding resolution. Another item on the list might be assessing whether the construction type listed is correct. You can query your database to assess whether the construction type is something other than unknown, but if it is how do you know if the data is current or accurate? When assessing data quality, it is important to profile the value of the data attributes, and if the data correctly describe the risk. Exposure data quality is also impacted by the processes involved in creation of the data.

We introduced data *value* profiling in Unit 2. In contrast, profiling data *quality* refers to the following processes:

- Identify the exposure data for quality profiling based on peril, region, and market segment considerations.
- Measure or quantify the potential data quality issues against business rule tests. Data value profiles, data queries, and data audits are used to provide measures.
- Analyze the nature and/or cause of the data quality issues using audits of processes impacting data as well as analysis interpretation.
- Improve the quality of data by focusing on the data improvements that provide the greatest return on investment.

FIGURE 5: Steps in Profiling Data Quality

When Data ≠ Exposure

Information is what you want based on the data you have. Ultimately what is important about data or information is whether or not those data are a true representation of what they describe. The question is therefore - *Does your exposure data represent your true exposure?* Finding out how well your exposure data represent reality is the core of data quality assessment.

Discovering *why* your data quality is poor is a critical aspect in changing the processes that impact your exposure data.

Listed below are the four primary reasons why exposure data may not reflect actual exposure:

1. **Incorrect data** are used to describe the risk. As an example, assigning limits instead of values to the coverage value field in your EDM will underestimate the value of your risk.

2. **Missing data** covers not only the lack of individual parameters such as year of construction, but also the potential for entire accounts/policies to not be entered. These missing policies could represent non-modeled regional perils. It may also represent data that are not captured, such as other business units of an insurance company that may have exposure to the same events.

3. **Misinterpretation of data or incorrect assumptions** can account for much of the discrepancy between data and actual exposures.

4. **Exposure data analysis errors**, many of which have been explored earlier in this course, may lead to incorrect data. Incorrect aggregation of data is one example.

The issue then is how to comprehensively identify and assess data quality issues.

Exposure Data Quality: Fitness for Purpose Framework

The “fitness for purpose” definition requires a structured analysis of data quality in a framework that will reveal problems and their root causes. A definition of exposure data quality might be: “Exposure data quality is a measure of the accuracy, resolution, and completeness of exposure information for catastrophe loss modeling as well as other exposure and capacity management purposes.” When assessing data/information quality for the product or service you are providing, you will first identify the following:

1. Data “fitness for use” and conformity to specifications for the desired analysis
2. Whether data meet the needs of the end-user
3. Data quality relative to…
    a. …the peril. Be it earthquake, windstorm, terrorism, or something else, each peril has specific exposure elements that have a greater impact on losses than others and are important to review.
    b. …the regional hazard level. For the peril in question, is high, moderate, or low?
    c. …the market segment. As with the perils, the relative importance of exposure characteristics can change between single family personal lines and excess policies on a large commercial risk

When assessing data quality, it is critical to understand the underlying regional issues such as what data are typically captured and available, the exposure resolution, and any common practices that influence key data characteristics.

Exposure Data Quality: How Imperfect is it?

In a perfect catastrophe risk analysis world, it would be possible to measure the quality of your data and formally quantify the differences from actual exposure. If this were the case, of course, one could presumably replace the incorrect or missing values with correct and accurate information. Instead, various proxies and heuristics are used to *estimate* the quality of a given set of exposure data and decisions must be made regarding what to do next, if anything.

Prior to discussing specific data quality measures, however, several complicating factors need to be considered in the assessment of exposure data quality. As an example, good data quality is defined and measured differently for different market segments, and data quality concerns may also be ranked differently across each segment. A common end goal is to rank data quality based on its impact on both modeled and actual losses. This is a difficult task, especially for less-understood perils. We will also discuss data quality in the context of the use of those quality metrics. If the data are being used for more than one purpose, they may be of sufficient quality for one but not for another. Discussing the quality of the data quality may seem a bit redundant, but best practices require that a data quality assessment be transparent and reproducible for it to be effective and trustworthy.

With this background, we will now delve into the details of measuring exposure data quality.

Measurable Impacts of Exposure Data Quality

While data quality has an impact on the accuracy of model results and the business decisions made from them, a measure of the impact is not so easily obtained. How data quality can be measured, and how that measure can be used as a competitive advantage in the marketplace, is a common catastrophe risk management theme. Data quality impacts modeled loss estimates, catastrophe event response, pricing, and claims versus modeled loss comparison, to name a few. Some data quality concerns that can be answered by their measurable impact on modeled losses as well as comparisons between modeled and claims loss include:

- What is the return on investment (ROI) of obtaining the best (or better) quality data?
- How does data quality affect company decisions? Which issues are the most business-critical?
- How can the data quality be improved?
- How accurate do the data need to be?

A considerable competitive advantage can be gained from better exposure data and hence better catastrophe model results. These downstream impacts include more accurate pricing, improved risk transfer decisions, and improved risk diversification decisions. These apply not only to internal applications, but also to interactions with counterparties that recognize the importance of data quality governance.

Methods to Assess Data Quality

Several methods can be used to measure data quality; however, the most common approach is to create **data value profiles** of key parameters (e.g. geocoding resolution, construction type, and related characteristics). These data queries are an appropriate approach to identifying the resolution and completeness of data values and can be performed frequently. In addition, these queries can be used to flag potential data issues that require further investigation.

Data quality audits are used to identify issues that are not easily assessed from data value profiles. These audits put a business framework around the categorization of data quality, which requires some interpretation of data value profiles. Analyzing a questionnaire filled out by all the groups or individuals that impact data quality is one approach to subjectively measuring the impact of the more intangible data components and processes.

Sophisticated software tools offer pattern matching (identify identical data patterns to cleanse or validate data) and smart clustering (similar data elements are grouped into a meaningful hierarchical pattern to identify data quality relationships). These types of analyses can be powerful tools in providing a measure of data quality issues. The RMS Data Quality Toolkit is another application that evaluates exposure data quality in a systematic and transparent way.

Categorizing Data Quality

A structured approach to assessing something as complicated as the quality of exposure data begins with categorizing its components. Data categorization, whether for data quality or other types of assessments, provides a consistent and structured approach that can be used in any peril, hazard region, or line of business. This consistent view of exposure data quality gives the analyst a way to identify and address systemic data quality issues whether it is applied to construction characteristics, valuation, occupancy, or policy coding. It also provides a framework for making data quality comparisons from account to account, portfolio to portfolio, and market sector to market sector. Finally, it provides a structure for creating a measure of data quality that is based on common critical categories of issues specific to catastrophe exposure.

Categorizing data quality, however, is not straight forward. For example, a single data value can be accurate (i.e. valid and useful) for one specific purpose, time, and place, but not for another. Following are three examples:

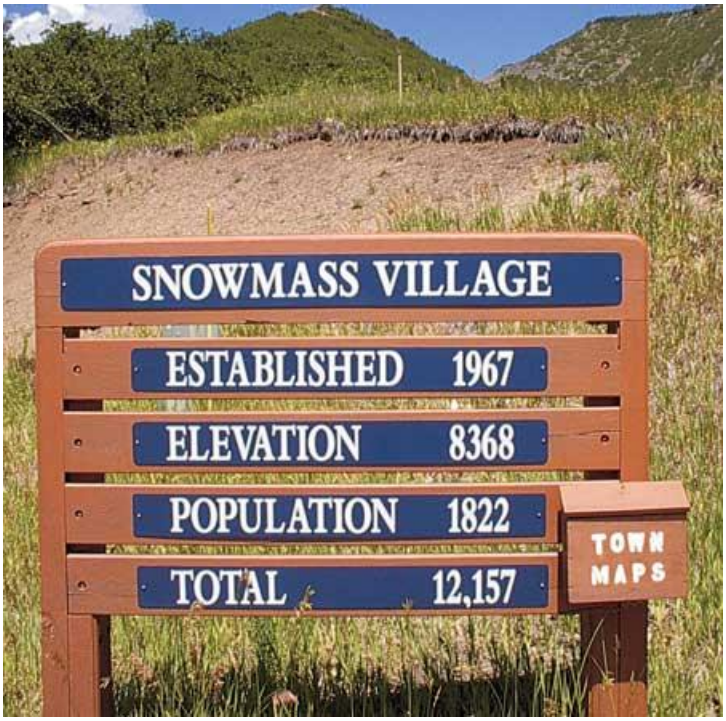
Example 1: Although capturing a secondary modifier of cripple wall may be a valid data capture, it is invalid if the primary construction class is steel moment frame.

Example 2: Capturing roof/wall connections is useful for U.S. hurricane analyses. However, while it may be a valid capture of data for this region, it may not be useful for a Europe windstorm analysis if the model being used does not consider this attribute.

Example 3: The location of a risk in the database was recorded two years ago at the postal code level. The address information that was valid then may no longer be *accurate* because the resolution of location information has improved to now include street address information.

Figure 6 provides a lighthearted example of a data quality issue. The data are all valid, but the interpretation of the values is inappropriate.

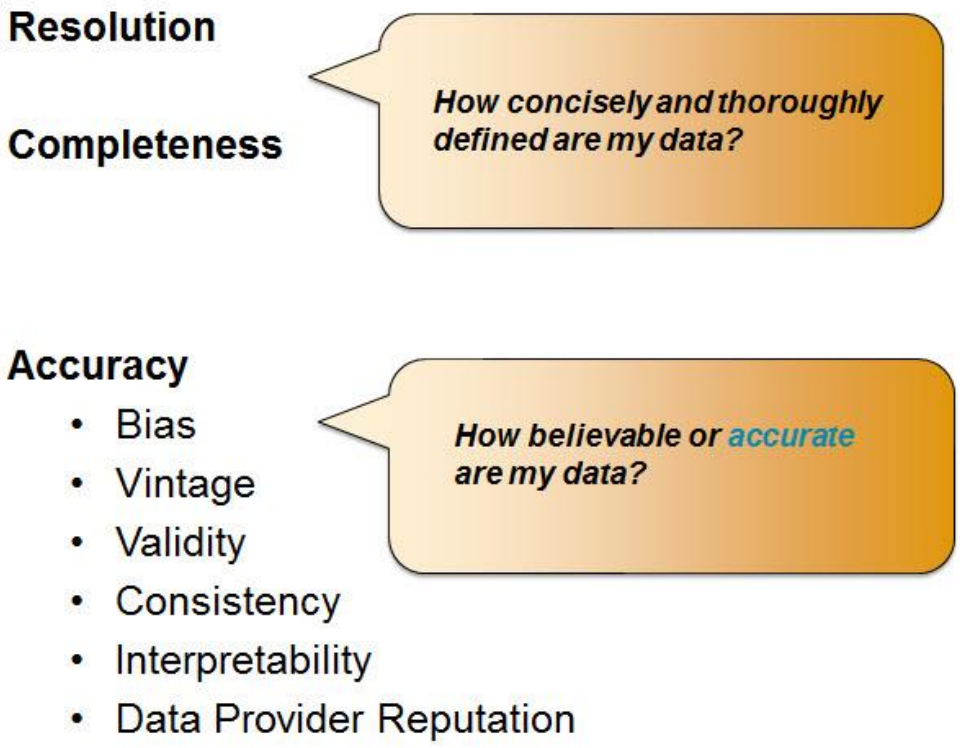
FIGURE 6: When Math Goes Wrong



Key Categories of Exposure Data Quality

The list of data quality categories that will be addressed for catastrophe modeling data are shown in Figure 7. Although the list is not exhaustive, these are generally considered the most important data quality concerns. Note that these categories are further grouped. The first two categories, resolution and completeness, describe how concisely and thoroughly the data is defined. The last category, accuracy, describes how believable or correct the data are and includes several sub-categories. *Note that some data quality accuracy issues can be characterized by more than one data quality category.*

FIGURE 7: Key Categories of Exposure Data Quality



The following sections will look at each of these data quality categories in more detail.

Resolution

Data resolution queries identify how thoroughly or concisely data are defined. This data quality category is easy to identify and measure via data value profiles. As an example, if you were asked to profile the geocoding resolution of a portfolio, you would arrive at a measure such as "95% of the portfolio limits are geocoded to street level resolution or higher." Another example is "35% of portfolio total insured values have construction type values resolved to at least two levels of resolution (e.g. RMS Construction Code 3A [reinforced concrete] or 3A3 [unreinforced masonry infill])." Yet another example of lower data resolution is capturing only two significant digits for a latitude / longitude coordinate instead of five (e.g. -24.29 / 151.63 vs. -24.29299 /

151.62971)[3]. It is important that both an identification of data resolution and a measure of its impact on the number of locations, total insured value, or limits, as examples, are included in the data value profile results.

Given that a high level of data resolution for some peril/region/market segments is less critical than for others, it begs the question “is obtaining the highest resolution and most complete data always the best goal?” The answer, as we discussed in Unit 1, is yes. Highly resolved data decreases the measured uncertainty, and increases the confidence in the data, data analysis results, and subsequent interpretation. In addition, the landscape of catastrophe risk analysis is continually changing, which requires a responsiveness and flexibility in the data management processes. In this environment, the more data that are collected up-front, the more likely one will be able to quickly take advantage of future modeling capabilities.

Completeness

Evaluation of data completeness identifies gaps in data and assesses the impact of these unknown data. Similar to data resolution, data completeness is easy to identify and measure via data value profiles. If you were asked to profile the completeness of a portfolio, you could query the data to determine the amount of missing primary vulnerability attributes (occupancy, construction, year built, and number of stories). For example, it may be determined through the data profile that construction class is missing (or set to unknown) in 25% of the portfolio locations, or that no building heights are specified at all. In addition, it is possible to determine if policies from a specific business unit have not been included in the portfolio. Both an identification of unknown data values and a measure of their contribution to the number of locations, total insured value, or limits, should be included in the data value profile results. Quantifying missing business can be more of a challenge.

The level of data completeness required depends on the purpose of the analysis. For example, if you are running an accumulation analysis to identify areas of exposure concentration, having the number of stories data field complete may not be critical for your purpose. However, if you are preparing the data for a catastrophe model to estimate potential earthquake loss, then having the number of stories field complete is necessary for an accurate analysis.

As with data resolution, highly complete data decreases the measured uncertainty, and increases the confidence in the data, data analysis results, and subsequent interpretation. In addition, having complete data will allow the exposure to be used for analyses other than their originally intended purpose. Finally, more complete data allows for better and faster business decision-making without having to accommodate for data assumptions and missing data.

[3] Depending on latitude, 0.01 degree is approximately 1 km, whereas 0.00001 degree is about 1 m. The number of significant digits should be appropriate to the positioning tool as well as the location.

Accuracy

The final data quality category is accuracy. Data accuracy identifies potential problems with the data. Data value queries can be used in the context of identifying the believability and credibility of the data, and to identify potential problems. However, data accuracy is more difficult to identify and is often not amenable to being measured via data value profiles. These queries serve as data quality "flags" of issues that require further investigation via an audit process. Some examples of data accuracy are:

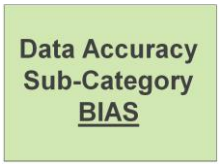
1) A South Carolina portfolio with some all-risk and multi-peril policies is being reviewed for data fitness for purpose for each peril. The data may be highly resolved and complete, but the policy terms as coded may be valid for earthquake but neither flood nor hurricane risk. Similarly, the provider of your data may be very knowledgeable about windstorm construction practices, but not earthquake. As a result, the earthquake sub-limits may have been entered as the limits in the wind portfolio in error.

2) You receive a portfolio and create data value profiles which show that a) only location coordinates are provided without street address information, b) all construction type is set to unknown but secondary modifiers have values entered, and c) all year built entries are entered as multiples of five (e.g. 1920, 1965, 1970, 1975, 1980, etc.) and thus appear to be biased. The queries highlight potential issues, and if there is no further documentation or reference information, these data are suspect, and its quality unknown.

3) As an analyst for a reinsurer, you create data value profiles for a Florida hurricane residential portfolio that show roof type and square footage are not entered and the accompanying RDM you receive shows analysis results in Euros (EUR). In addition, the profiles show that the data contains 20-story buildings coded with a wood frame construction class. This highlights potential data validity problems for this peril and region.

The following section provides detailed examples of how the combination of data profiling and audit questions can provide insight into potential data accuracy issues. We will examine five of the data accuracy sub-categories: bias, vintage, validity, consistency, and interpretability.

Data Accuracy Sub-Category 1: Bias



Data bias can have a significant impact on all catastrophe risk management decisions based on exposure data analysis and model results. ***Data bias is the systematic prejudicial or preferential entry of data regardless of actual value***. Potential data bias issues are easily flagged by profiling if you are familiar with some of the ways bias can be introduced into key data characteristics.

Listed below are examples of data profiles that can flag potential data bias:

- Query the values for the primary data characteristics (construction type, year built, number of stories, and occupancy) to assess whether all or a vast majority are assigned the same value. For example, if all unknown construction types are defaulted to wood frame in Florida, the data are biased and probably conservative.
- Query secondary characteristic values to assess if the type of value entered is appropriate for the construction type, as well for which combinations of characteristics are being entered. For example, if the only secondary characteristics with values entered are Construction Quality and Roof/Wall Anchor, and the same value is consistently entered for all risks, then this may indicate preferential or “rote” entry of data.
- Query coverage values to ascertain whether they are consistently calculated as the same multiplier of some other data parameter. Are the contents values always a factor of 0.80 times the building value for a general commercial book of business? Follow-up interview questions should be asked to assess why and when the values are defaulted in order to identify if indeed it is the contents values that are being assumed, and if the actual values can be provided to improve data quality.

Below is a list of sample data bias checks and questions to consider when assessing and measuring bias:

- Are assumptions for unknown values *documented, justified, and verified*? If so, when were they last validated against current business practices and fitness for purpose? Assumptions may have been made that were appropriate at one point but no longer fit the current business environment.
- Is information processed and recorded from a restricted list of choices or from formats that may not reflect the actual range of data?
- Is there a business hazard (e.g. capacity and exposure constraints) influencing the representation of data?

- Are the data parameters defaulted in such a manner as to potentially bias values (e.g. default of year built field).

Data Accuracy Sub-Category 2: Vintage

Data Accuracy Sub-Category VINTAGE

Vintage is the believability, verification, or accuracy of the age of the exposure data, the database, or underlying information relative to the date on which is being analysed or reviewed. Assessing potential data vintage issues (verification of the age of data) can be flagged by profiling date, time, or status (e.g. policy status) entries in either the exposure database or the database transaction log.

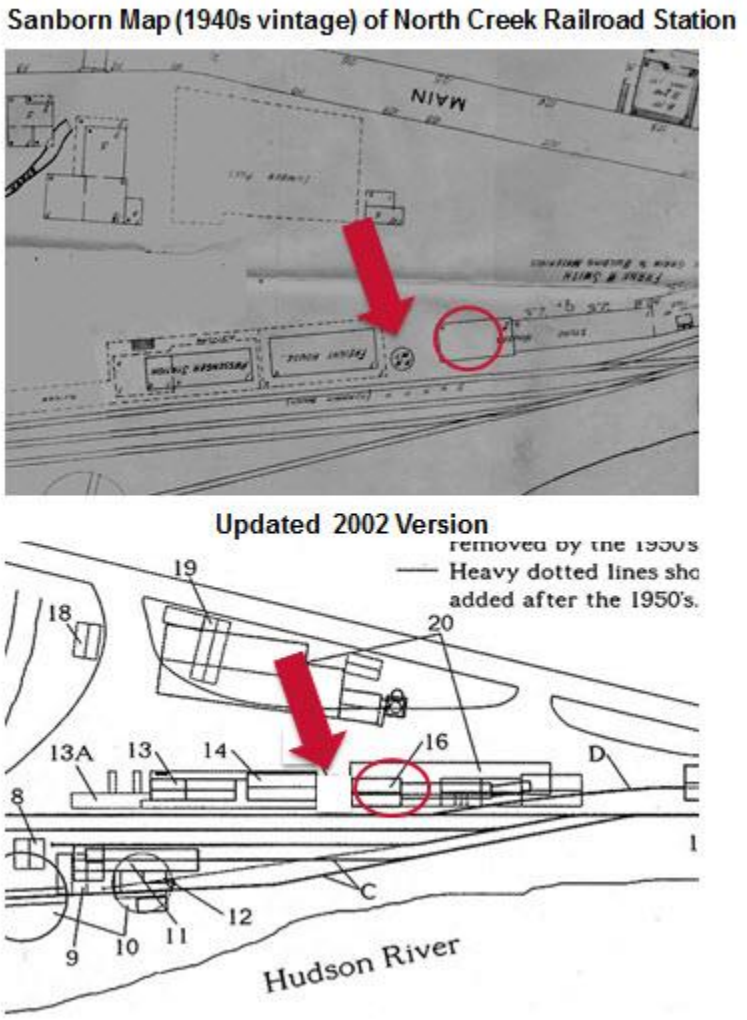
Listed below are examples of profiles to identify data vintage and flag potential vintage bias:

- Query source data or exposure database against date fields (e.g. policy inception and expiration dates) and compare to queries of final EDM date fields. Policies may be included that were in run-off and should no longer be on the books. In addition, policy expiration dates indicate if there are expired policies in the database.
- Review database properties for database creation, modification, and backup dates. Compare with underlying table creation and modification dates for consistency.
- Query policy status fields and compare to policy inception and expiration dates.

Out-of-date information can impact data quality. Figure 8 on the following page shows a Sanborn map of unknown vintage on the left. Sanborn maps have been used by underwriters and agents for decades to assess risk. A red circle highlights the Frank Smith stores (the building being insured). The red arrow identifies the location of a water tank next to it.

The map on the right is a 2002 map that shows the same red arrow identifying the insured building, but note that the red circle shows that the tank has been removed and that the store is no longer at risk to hazards from the water tank.

FIGURE 8: Example of Vintage Issue



The following is a list of sample data vintage checks and questions:

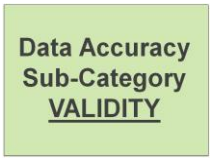
- Is the vintage of the database and underlying data recorded and appropriate for your purpose?
- Are the policy inception and expiration dates updated and checked against the submission or other policy information sources?
- Are policy status flags used and updated whenever the inception and expiration dates are updated?

The business implications of data and database vintage can be significant. For example, insurers are faced with a potential over- or under-estimation of modeled losses if out-of-date information is used, depending on how significantly business has changed over time. Similarly,

old information has led to the incorrect assessment of real-time modeled losses to actual events for catastrophe response purposes and claims comparison. If insured values have not been updated in the exposure database to account for increases over time, the modeled damage ratios will not be applied to the proper replacement costs.

For reinsurers, the vintage of exposure data directly impacts the reliability of treaty pricing and baseline data comparisons. Complex reinsurance structures that change over time are another example where data vintage can have a significant impact.

Data Accuracy Sub-Category 3: Validity



Validity is the verification of data to be valid and correct against various outside resources such as field checks, alternative data sources (i.e. industry data), and company processes. Assessing potential data validity issues (verification that data is valid and correct) can be flagged by profiling values to find outliers or those critical fields with no data or suspect information.

The following lists examples of profiles to flag potential data validity issues:

- Query the number of buildings field for suspected aggregate locations.
- Query fields with numerical values to flag those that are much higher or lower than the average value. For example, you may find potential data issues when querying construction class and building height together. The classic example is a wood frame building coded with a building height of 50 stories (not realistic).
- Query fields with numerical values to see if they compare as expected with past values.
- Query construction type and compare with industry data to find outliers to trends

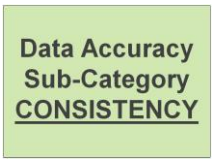
Suppose you run a data profile on construction type, and discover that there is a large percentage of residential earthquake policies in northern California coded as having masonry construction. Based on information from a residential lines industry exposure database, a much smaller percentage of homes in this region are masonry. This discrepancy from the industry average raises a flag that requires further investigation. Upon request, you obtain photographic proof of the building construction. A preliminary assessment indicates the masonry construction designation is correct. However, further on-site inspection reports show the masonry-coded structures to be primarily wood frame with a masonry veneer cladding. This example highlights that often more than one type of questioning or evidence needs to be reviewed to fully assess data validity.

The following is a list of sample data validity checks and questions:

- Are field checks made for the location of risks, key building characteristics, coverage values, and secondary construction modifiers?
- Are data compared to independent data sources, such as the RMS ExposureSource databases?
- Is individual risk information retained in a separate database if aggregated? Is number of buildings accurately recorded?

The business implications and impact of invalid data on the accuracy of modeled loss results and subsequent business decisions is large. Often, the necessary process changes are easy to implement once all the data stewards and owners understand the issues and downstream impacts of their decisions.

Data Accuracy Sub-Category 4: Consistency



Consistency is ensuring that the data format and data definitions are uniform and reliable from record to record. Assessing potential data consistency issues can be flagged by profiling data in fields for consistency in formatting and data entry.

Following are examples of data profiles to flag potential data consistency issues:

- Query coverage values to assess if they are always entered separately for building, contents, and time element coverages.
- Query limits to assess if they are always entered or only entered when different from values.
- Query deductible field to assess whether deductibles are consistently entered as monetary amounts or percent of values.
- Query construction and occupancy schemes/types to see if a single scheme is used throughout the dataset.
- Query the street address field to see if the street number and address fields are entered in a consistent manner.

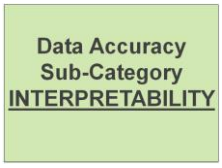
Although queries may show consistent entry of street number and address fields, further questions and checks should be made to determine how information is entered if there is a cross-street designation or if the street address is not known or complete. Other information may need to be checked to understand the impact of data entry rules, which may result in variable data entry practices. An example is the inclusion of postal code information unless the city is known.

The following is a list of sample data consistency checks and questions:

- Is the location address entered the same way for every record within each country?
- Are there multiple construction class schemas entered for the construction type?
- Are there multiple occupancy class schemas entered for the occupancy type?
- Are questionable or unknown data treated the same throughout the dataset (e.g. defaulted, left blank, flagged, etc.)?

Data consistency can have a significant impact on all catastrophe risk management decisions based on exposure data analysis and model results. It is one of the more difficult data quality categories to characterize, and it is even harder to assess its impact on exposure data analysis results. Understanding consistency requires rigorous auditing to identify potential issues since data profiles alone are not adequate for flagging potential issues.

Data Accuracy Sub-Category 5: Interpretability



Interpretability is the clear and concise representation of data so that they can be reliably understood through supporting documentation. Potential data interpretability issues can be identified by profiling data comment fields and flagging potentially incomplete field entry.

The following are examples of data profiles to uncover potential interpretability issues:

- Query for entries in record comment fields.
- Query for entries in other text fields that are not used for modeling.
- Query unique data identifiers for non-unique values.

For example, if construction codes are included in the exposure data, it is necessary to have the definitions of these codes in order to properly interpret what they mean. If the definitions are not included, or if they are incorrect, this can result in errors in both exposure analyses and catastrophe risk analyses. Referencing data format documentation or data dictionaries would be critical prior to improving the quality of these data.

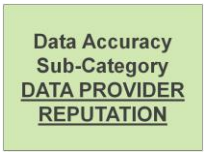
The following is a list of sample data interpretability checks and questions:

- Do data format and representation guidelines and rules exist?
- Are there data definitions and format documents?
- Is descriptive information on the data attributes included in a data dictionary?

- Are the data derived from paper records and can they be reliably read or interpreted?

As with all the other data quality categories, data interpretability issues can have a significant impact on all catastrophe risk management decisions that are based on exposure data analysis and model results. Similar to data consistency, it is one of the more difficult data quality categories to characterize, much less assess the impact on exposure data analysis results. It requires rigorous auditing to identify potential issues since data profiles are not always adequate in flagging potential interpretability issues.

Data Accuracy Sub-Category 6: Data Provider Reputation



Data provider reputation refers to whether or not you know and trust the source of your data. The data source could be an individual who is recording the building characteristics of a property you insure while on site. The source could also be the person who is entering those building characteristics into a database to store the information, or the person who is extracting it from the data for your use. All of these sources come together to influence the reputation of the data providers and its sources.

Now that you have an understanding of the data quality categories, please review the following e-learning interaction. This interaction will give you an opportunity to identify data quality issues given a set of examples and will reinforce the data quality category definitions.

E-LEARNING INTERACTION 2: Exposure Data Quality

To begin the interactive review of the key categories of exposure data quality, go to the Exposure Data Analysis course in the CCRA portal of Owl and select Interaction 2: Data Quality.

Data Resolution vs. Data Accuracy

In the context of the preceding section on data quality categories, you should now have a good understanding of data resolution, data completeness, and data accuracy. It is important to stress that **high data resolution does not imply high data accuracy**, **nor does low data accuracy imply low data resolution**. This is why data value profiles or queries alone do not provide the full data quality picture.

Large regional differences in data resolution and accuracy can exist for the same peril, both in the context of what exposure data information is available or captured and in the context of what is required for modeling. This may be due to different insurance regulatory environments,

different regional business practices, or differences in education about the characterization of exposed risks. Conversely, if the same exposure data set is used to characterize multi-risk exposures, it is possible that the resolution or accuracy of the data is of sufficient quality for one risk, but not for the other.

It is helpful to provide a few examples in the context of geocoding to highlight the differences between the pieces of exposure data quality assessment. Geocoding resolution provides information on the level of geocoding match assigned to a location. A data profile on geocoding resolution will provide you with a measure of the percent of locations, values or limits that have a specified geocoding level. The resolution measure will not, however, provide information on the accuracy or believability of that information. Geocoding accuracy provides an assessment of which geocoding data is correct within the context of analyzing catastrophic risk to a particular structure. For example, an office building's address is listed as 10 Main Street and is assigned a high-resolution geocoding match. However, the address might be for the home office, not the satellite office being covered by the policy, making the data inaccurate.

Data Quality Audits

Identifying data accuracy issues can lead to identifying process-oriented data quality issues. Internal audits that develop a better understanding of the processes that impact data quality are critical to providing a full picture of the accuracy or believability of your data. Unlike data profiling where queries can be run on your data frequently, data audits are usually viewed as an annual project to identify and improve data quality issues. Subsequent data quality audits would then be compared against past audits. Because a structured approach would be used, it would be possible to make data accuracy comparisons with other books of business.

Using data audits to examine the impact of exposure data generation processes on the impact of data quality usually requires on-site visits to conduct interviews. Interview questionnaires may be tailored to provide further insight into potential accuracy issues flagged by data profiles and other data analyses. In order to cover all potential processes, interviews are ideally attended by anyone in the organization that "touches" exposure data and exposure data analysis results, including systems and IT groups, underwriters, data providers, analysts, and decision makers. Very importantly, these interviews should be consistent across business units and companies.

To reiterate a critical message mentioned in nearly every unit of this course, data quality issues are conditional on line of business, hazard region, and peril. Exposure data quality identification is also conditional on product or policy requirements (e.g. how the exposure data are provided and by whom) and the data format or platform (e.g. paper submission, digital format, or direct entry into EDM).

Analyzing Data Quality Measures and Audits

Once the data quality measure is complete, the results need to be examined and interpreted to determine potential data and process issues. Depending on the fitness for purpose of the data, data may be deemed high quality for one use, but not another. So when analyzing the data quality it is especially important to keep the purpose in mind.

Data quality measures can highlight potential issues with the data, primarily with respect to the data resolution and completeness. The data accuracy component is more difficult to analyze without further investigation in the form of data audits.

Finally, once the identified data measure is interpreted, the cause of the data or process issue can be identified, and thus improved upon.

Improving Exposure Data Quality

Once data quality issues are identified and diagnosed, the next step is to implement improvements on both the processes and data that will improve overall exposure data quality. Improving the acquisition, processing, storage, and retrieval of data can all have positive impacts on data quality.

Approaches to improving the data include:

- **Data Cleansing** – This includes fixing data problems and verifying data. An example would be the cross-checking of building valuation against other valuation products.
- **Data Integrity Improvement and Enhancement** – This includes appending or inserting external data to existing data sets to enhance or increase the integrity/value of the database. This can include inserting building square footage for residential hurricane locations, or appending secondary construction characteristics from another source of information.
- **Data Migration and Merging** – This is the process of combining several databases to create a single record with links to data. This process is often regarded as the creation of metadata, which improves data quality through interpretability if it provides a link to the source of data or information. Companies often merge data from different databases together to create a single, larger database. An example would be the combining of location data, policy data, and reinsurance data to create a “master” exposure database such as an EDM.

For most time- and resource-constrained companies, the goal is to implement data improvements that provide the greatest return on their investment. This requires understanding the data issues that have the greatest impact on key metrics, such as modeled loss vs. actual loss or modeled loss uncertainty measures. These include:

- **Inserting higher quality data from other databases**. If the geographic distribution and make-up of your business can be compared to an industry-wide database, then this information can be used to verify existing data and update assumed or missing information.
- **Training and education** for upstream users of data will improve the data products and the implementation of processes used in their creation.
- **Creating and enforcing data generation, manipulation, and analysis processes** will also improve data quality. A data accuracy audit may highlight the need for on-site inspections in certain regions and documentation for how and when default assumptions are made.
- **Development of "smart systems"** that consistently and automatically use rules and formulas to enter data can also be used. While smart systems are useful tools, they have the potential to cause data bias and become outdated if not routinely updated.

Note that this is not a comprehensive list of data improvement methods.

One approach to identifying where data quality improvements should be focused is to perform an internal audit of data quality following significant catastrophic events. As an example, the top 50 account claim files can be pulled, which will likely represent an appreciable percentage of the event loss. As these claim files are reviewed the exposure data should be checked against the specifics of the exposure in the claim file to verify address information, missing locations, and correct policy terms as compared to those applied by the adjuster. In addition, accurate key modeling parameters, non-modeled loss contributors, and any secondary modifiers that could have been coded to more accurately characterize the risk in question should be identified. Going through this process will likely result in identification of data quality weakness trends that can be addressed in order to fine tune the data capture process.

Why Does Exposure Data Quality Really Matter?

The quality of exposure data has been found to have a significant impact on modeled losses for all perils, regions, and market segments. In addition, exposure data quality can have an impact on the measured uncertainty around modeled loss results. The modeled loss uncertainty measures are directly impacted by location geocoding resolution and interpolation, uncertainty in construction characteristics, and coding of known secondary modifiers. This topic is discussed in detail in the Uncertainty Measures course.

Coding of financial information has been covered throughout in this course, however, it is helpful to summarize that underinsurance, calculation of aggregate values, incorrectly coding limits in place of values, and miscoding of complex policy and treaty structures are common data quality issues. The coding of business interruption (BI) values is one aspect of financial

coding for which there is no single best practice approach across all peril regions and market segments. Estimating business coverage as a monetary amount over a specified period of time is not straight forward. Unfortunately, this results in the lack of data capture for this financial parameter in many businesses. During the 2004 U.S. hurricane season, it was discovered that 5% of total insured losses for BOP (Business Owners Policy) insurance was due to BI coverage that was never captured in the exposure data.

Location resolution is covered in detail in the geocoding and hazard retrieval course. It is important to highlight that not only the resolution of the location is important, but also the accuracy. It is not uncommon to find insured structures that have been assigned addresses for home office or mailing address locations far from the actual site.

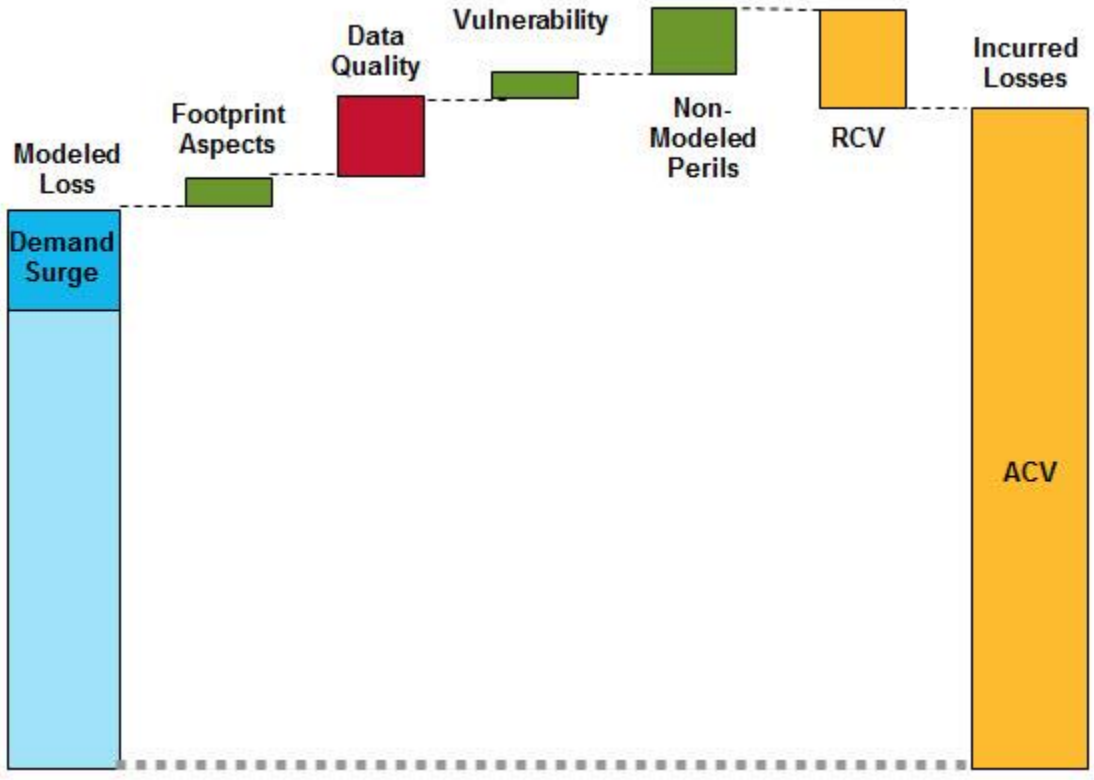
Primary characteristics or key building characteristics are discussed in great detail in the RMS CCRA Training Program peril model courses. An example highlighting the impact of incorrect coding of construction class is the coding of light metal instead of steel frame construction for commercial books. This can result in double-digit loss differences.

Data quality is a significant contributor to differences between modeled estimates and actual loss, commonly accounting for 10-15% or more of actual loss that is absent from the modeled estimates. Contributing data quality factors include the vintage of exposure databases used in comparisons, incorrect insurance to value representation (underinsurance), incorrect coding of data, missing exposure, geocoding resolution, mapping of schemas, data assumptions, and default values.

Figure 9 shows a claims analysis from the 2004-2005 U.S. hurricane season. Claims were analyzed to characterize the differences between modeled loss results and actual losses. This diagram is an illustration of the relativities between various components contributing to the loss differences and is discussed in detail in the Tropical Cyclone peril course. You will see that data quality is a close second to replacement cost value (note: observed claims data is often reported as actual cash value [ACV] and the replacement costs [RC] are on average 20% higher). In general, it was found that the following data quality issues could be identified along with their associated contribution to loss:

- Vintage of exposure database: < 3%,
- Insurance to value (underinsurance): 10-15%
- Coding of data (construction class, number of stories, financial coding): ~0-5%.
- Other data quality issues that were identified include: missing exposure, geocoding resolution, mapping of schemas, data assumptions, and defaults values.

FIGURE 9: Claims Analysis 2004-2005 U.S. Hurricane Season



Market segment differences were also identified, and data quality was a significant contributor to loss differences in personal lines, commercial lines, and excess and surplus (E&S) lines. For E&S writers, a considerable challenge is managing the vast amounts of data that are funneled into a modeling platform in a very short period of time. This situation results in data entry assumptions and defaults for key parameters. For this reason, data quality is a significant contributor to the gap between modeled and actual losses.

Exposure Data Quality Issues in the Market

Cleansing data, and setting in place processes that ensure and monitor data quality, is expensive and time-consuming. Knowing that the data quality for specific location characteristics may matter more in some peril/regions than others begs the question: Is obtaining and maintaining the highest data quality always the best or most reasonable goal?

The answer is a resounding “YES!” One must also acknowledge that there are business realities that require incorporating a cost/benefit analysis in the decision-making process. Numerous catastrophic events have highlighted the benefit of high data quality, thus driving companies to spend more money now to better understand their losses for future events.

Some benefits of monitoring and ensuring high quality data include a decrease in measurable uncertainty in analysis results, an increase in confidence in business transactions (e.g. pricing), and streamlined structuring of data acquisition/processing efforts.

Meanwhile, cost considerations may include proactive cost avoidance (decision to spend money now vs. later) and optimizing a trade-off between business costs and data quality costs (focusing efforts on data quality issues that have the greatest impact to modeled losses vs. all issues at once).

Increasing regulatory and standards requirements are also driving the adoption of higher data quality standards. These include U.S. DOI reviews, regulatory agency ratings, and the Gramm-Leach-Bliley Act (GBLA). In addition, Solvency II for the EU imposed data quality standards as part of meeting economic capital requirements.

The hard vs. soft market cycles have also driven the adoption of higher data quality standards. Harder markets make it easier to push for better data quality. Furthermore, the recent era of instantaneous access to communication and knowledge means that up-front confidence in data quality can no longer be ignored or compromised. A company that responds more quickly may have the competitive advantage, especially if the quality of their data is high.

It is important to understand that the landscape is continually changing in terms of data quality best practices and acceptance levels. Pressure applied consistently upstream on the flow of exposure data can quickly change the risk analysis and data quality landscape. As more refined claims and exposure data are utilized by modeling firms to update and upgrade catastrophe models, the differentiation of risk between key modeling parameters is becoming more pronounced. This is due to a migration away from using an engineering judgment approach to fill in information when there is a lack of claims data, towards quantification of the risk differentiation via observable data. The key point here is that the relative importance of capturing key modeling parameters is dynamic. Staying current with model upgrades remains critical when determining data quality priorities.

Data Quality Identification and Improvements

The quality of exposure data has become a major concern for many insurance and reinsurance companies. Catastrophe models have become very sensitive to data quality as they have become more sophisticated. The quality of exposure data not only impacts the mean loss amount estimated by the models, but also the uncertainty in the loss results themselves. Poor quality exposure data results in unreliable catastrophe model output. Data quality is a controllable element of uncertainty in catastrophe models.

Insurance companies who focus on data quality are able to differentiate themselves with reinsurance companies, and even among rating agencies. Not only does better data lead to more accurate pricing and reinsurance placement, but it can also lead to premium credits for insurers as they purchase reinsurance. Insurers and reinsurers need high quality exposure data to demonstrate robust risk management practices.

Data quality in the industry has been steadily improving since the 2004-2005 U.S. hurricane season when insurers and reinsurers were surprised by the magnitude of their actual losses compared to modeled estimates. As discussed, many insurers we surprised to learn that data issues were a significant driver of those differences. Despite progress, however, the quality of

data describing catastrophe exposures still remains problematic. In particular, there is significant variability in the quality of the exposure data from account to account, underwriter to underwriter, portfolio to portfolio, and even within companies as well as across the industry. Because of this, there is a need for a standard metric that can quantify the quality of exposure data from one portfolio to another.

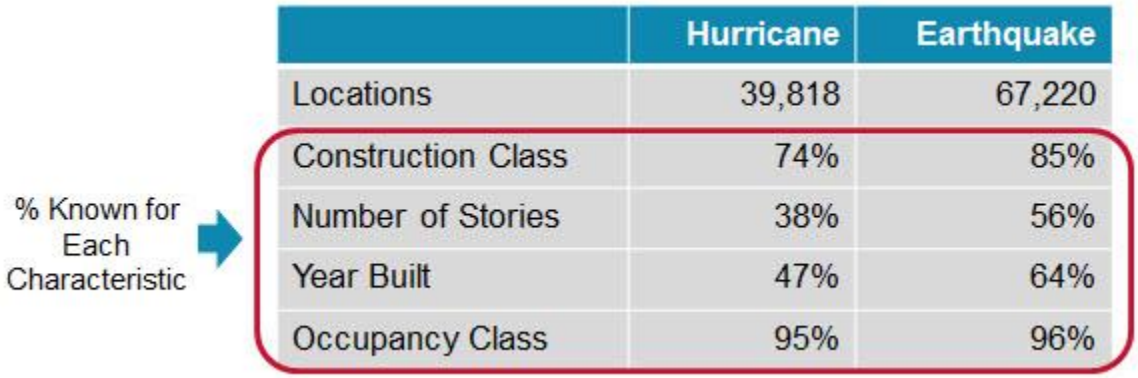
While it might seem that it should be quite easy to quantify the quality of exposure data, it is actually very difficult to measure in a systematic way. When there is no standard, reproducible measure, it becomes a challenge to establish proper objectives and manage the desired outcomes, and to incorporate the metrics into business decisions. There are no established methods or standard measures of data quality in terms of model results.

Data profiles are a conventional method of assessing data quality; however, they only show part of the picture. While a data profile may provide information on the completeness and resolution of the data, it does not tell you anything about the accuracy of the data, nor does it measure the importance of each of the attributes with respect to the vulnerability of the peril. For example, a data profile may reveal that a hurricane portfolio has 85% of its location geocoded to a street address level, while an earthquake portfolio has only 52% geocoded to the street address level. If you only looked at this information, you might make the conclusion that the quality of the geocoding data is better for the hurricane portfolio than for the earthquake portfolio. While the hurricane portfolio is better geocoding resolution, it does not necessarily imply it is better quality.

However, what the profile did not tell you is where those high-resolution geocoded exposures are located. Let's say that the earthquake portfolio is concentrated in eastern California where the hazard is relatively uniform across postal codes. In that case, having postal code geocoding resolution may be adequate and appropriate for the analysis. Furthermore, perhaps the hurricane portfolio is all in the state of Texas, which has substantial inland exposure. Of the 85% that are geocoded to street level, how many are coastal exposures vs. inland exposures? Additional profiling is needed to better understand geocoding within the context of the exposed peril and geography.

Figure 10 shows these two portfolios profiled by the percent of primary construction characteristics which are known. Recall that the relative importance of each of these characteristics varies by both peril and region. For example, the number of stories is an important attribute for the earthquake peril and knowing that there are locations with an unknown number of stories is critical to evaluating the quality of the data. However, for the hurricane peril, only 38% of the locations have a known number of stories. Because this attribute has less of an impact on hurricane loss results, it is not as critical to fully capture this information as it is for earthquake. Of greater concern for hurricane is that only 47% of the locations have a known year built, which is an attribute that has a significant impact on hurricane losses.

FIGURE 10: Sample Statistics from RMS EDM



	Hurricane	Earthquake
Locations	39,818	67,220
Construction Class	74%	85%
Number of Stories	38%	56%
Year Built	47%	64%
Occupancy Class	95%	96%

RMS has developed a way of quantifying the quality of exposure data to account for the importance of individual attributes with respect to vulnerability to a particular peril, and to create data quality analytics that will provide a systematic, transparent way of measuring exposure data quality.

The RMS Data Quality Toolkit combines RMS metrics, heuristics, and a best-of-breed property database into a single desktop application to assess exposure data quality. The Toolkit identifies data quality issues that are the most important with respect to the peril and region. It compares relative data quality by cedant, account, or business unit, and also assesses data quality improvements over time. Suspicious data, such as bulk coding (bias), can be identified as well. The reports created during the data quality analysis can also be shared with the market in a standardized format to allow rating agencies and reinsurers to compare one company or portfolio to another in terms of data quality.

While further details on the Data Quality Toolkit on not part of the CCRA Training Program, we encourage you to view the information provided on RMS Owl to further enhance your knowledge of the product.

Unit 3: Closing Key Concepts

A “fitness for purpose” framework for data quality provides context for what is important. All exposure data quality must be assessed relative to the exposed peril, line of business, and relative hazard region.

Data completeness, resolution, and accuracy are three important data quality distinctions. In addition, the accuracy sub-categories of bias, vintage, validity, consistency, and interpretability are equally important. Exposure data quality can be categorized via these data accuracy categories allowing for the comparison of data accuracy between different portfolios or accounts. Data value queries and data quality audits are two of several approaches to identifying exposure data quality problems.

Data quality identification should include a consistent and structured approach applicable across a broad range of geographies, perils, market segments, and level of data resolution. Data quality impacts both modeled loss and loss uncertainty. Finally, data improvements should be focused on the area that will provide the greatest return on investment and have the greatest impact on analysis results.

Now that you have completed the reading for the Exposure Data Analysis course, please move on to the Exposure Data Analysis Exercise in the CCRA portal of Owl.

Once you have completed the exercise, you must take the online assessment for this course, Exposure Data Analysis 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 Exposure Data Analysis course have been completed.