Exam 4: Introduction to Catastrophe Risk Management
Learning Objective No. 27-34
Study Note

The purposes of this study note is to be used as a guide for Learning Objectives 27-34 which cover the topic of determining and testing data quality

- LO 27: List the most common exposure data fields and their purpose. COPE data, primary and secondary modifiers.
- LO 28: Describe aggregate data, potential pitfalls of using this type of data & possible methods of disaggregation.
- LO 29: Determine and explain the impact of poor address accuracy and geocoding on modeled loss, understanding the materiality differs by peril and region.
- LO 30: Describe instances where mailing address may have been provided in place of risk address.
- LO 31: Explain the difficulty of obtaining accurate address information for certain classes of business (such as builders' risk, offshore energy, workers compensation).
- LO 32: Define bulk coding along with its potential uses and pitfalls in modeling.
- Lo 33: Identify validation checks which would indicate potential erroneous data or poor data quality. (e.g. 100 story wood frame building)
- LO 34: Explain the rationale behind assessing data quality and the effect it may have on the results. Peril specific examples.

This document contains content from the following documents and websites, along with original content.

- Natural Catastrophe Risk Management and Modelling: A Practitioner's Guide: 1.9.1.3 Primary and secondary modifiers
- COPE: https://www.investopedia.com/terms/c/cope-insurance.asp
- Bulk coding (Uncertainty in exposure data section): http://www.lmalloyds.com/AsiCommon/Controls/BSA/Downloader.aspx?iDocumentStorageKey=cc44f6be-b83f-4cf9-903c-e802c1f312a8&iFileTypeCode=PDF&iFileName=Understanding%20uncertainty%20in%20cat%20modelling%20for%20non-cat%20modellers
- 7 Step Quality Process: https://www.moodysanalytics.com/risk-perspectives-magazine/managing-insurance-risk/insurance-regulatory-spotlight/data-quality-is-the-biggest-challenge
- Geocoding Resolution: https://www.willistowerswatson.com/en-US/Insights/2018/05/geocoding-the-underappreciated-science-of-catastrophe-modeling

---

Learning Objective 27: List the most common exposure data fields and their purpose. COPE data, primary and secondary modifiers.

The following resources contain information on exposure data including location information, COPE, primary and secondary risk characteristics and financial information.

The Natural Catastrophe Risk Management and Modelling: A Practitioner's Guide Section 1.9. provides information on location data fields

COPE: https://www.investopedia.com/terms/c/cope-insurance.asp

Learning Objective 28: Describe aggregate data, potential pitfalls of using this type of data & possible methods of disaggregation.

In terms of catastrophe modeling exposure, aggregate data refers to information provided on a grouped level. Common aggregate levels include postal code, county, Cresta, or similar. This level of exposure data is in contrast to coordinate or address level detail.

Using aggregated data can have significant impacts on modeled loss estimates due to a number of factors:

- Correlation between buildings may not be properly represented
- Financial terms such as limits and deductibles may not be properly captured causing gross loss calculations to be incorrect. This can be particularly troublesome with commercial and multi-location policies that cross territories.
- Damage calculations for perils which are highly sensitive to location information could be incorrect by either missing potential loss scenarios or calculating a too large damage amount or accumulation. Perils that are highly sensitive to location include Severe Storm, Flood, Wildfire, and Terrorism.
- Not all perils and regions are built with sensitivity to specific location or address detail. It is important to understand the methodology for developing the model, peril and region of interest prior to spending resource where aggregate data may not be impactful.

Depending on the line of business there are a few disaggregation techniques that can be used to disaggregate the information. Some models have an analysis option available for this purpose. US Census data can be used as a guideline for spreading personal lines information across a more refined territory based on population. Modeling firms also develop industry exposure databases which can be used for multiple lines of business similar to the Census data. Additionally, there are companies specializing in data capture and data quality that may have additional tools available.

Learning Objective 29: Determine and explain the impact of poor address accuracy and geocoding on modeled loss, understanding the materiality differs by peril and region.

As with aggregate information, some regions and perils are highly sensitive to risk location. Poor address information and geocoding can lead to erroneous loss information. Locations may be placed where hazard is higher or lower. Additionally, locations may be placed where other portfolio locations exist leading to larger than expected loss or accumulation in an area and/or missing geographic regions where loss potential does exist.

Geocoding Resolution: https://www.willistowerswatson.com/en-US/Insights/2018/05/geocoding-the-underappreciated-science-of-catastrophe-modeling

Learning Objective 30: Describe instances where mailing address may have been provided in place of risk address.

Workers' Compensation, Accident& Health, Life risks and multi-location policies are all at risk for not capturing addresses where the risks are located but instead containing the mailing address of the main policyholder or headquarters. Implication of this mailing address is similar to poor geocoding and aggregate data described above.

Learning Objective 31: Explain the difficulty of obtaining accurate address information for certain classes of business (such as builders' risk, offshore energy, workers compensation).

Accurate address information is more difficult to obtain for specific classes of business in which there may be large schedules of risks under one policy or the people for a risk may be dispersed. Some examples of this include:

- Contracting business where the number of employees at the insured location may be inaccurate due to the contractor's location at the job site.
- Builders' Risk and Inland Marine business may insure equipment that is moved between job sites and often not at the insured's location.
- Company systems may not be capturing latitude/ longitude coordinates for Offshore Energy platforms
- Mobile risks, such as marine and auto, make it difficult to pinpoint where the values are at the time of the event.
- Property and workers compensation policies insuring a campus style location often just capture the administrative building but not all buildings on the campus.

Learning Objective 32: Define bulk coding along with its potential uses and pitfalls in modeling.

Understanding Uncertainty in Catastrophe Modelling For Non-Catastrophe Modellers coding (Uncertainty in exposure data section): https://www.lmalloyds.com/cmg

http://www.lmalloyds.com/AsiCommon/Controls/BSA/Downloader.aspx?iDocumentStorageKey=cc44f6be-b83f-4cf9-903c-e802c1f312a8&iFileTypeCode=PDF&iFileName=Understanding%20uncertainty%20in%20cat%20modelling%20for%20non-cat%20modellers

The term bulk coding represents making a universal assumption across a portfolio or line of business in order to allow cat modelling to proceed despite missing information. This assumption can be made either by the person preparing the data for modelling or by the cat model itself.

Systematic data issues are often the cause for missing information. These systematic issues can lead to bulk coding and can result in biases across a portfolio. Further impacts are losses being consistently under- or over-estimated. For example if Year Built is not provided on a book that is actually designed around providing cover to older or newer properties, then the assumptions used will likely be consistently outside of the actual year built.

While data quality assessments often check for instances of bulk coding, not all bulk coding is impactful on loss estimates. For example, assuming the number of stories of “2” for wood frame, single family dwellings within a homeowner’s portfolio will likely have minimal impact on losses as this is a highly probable characteristic of this construction and occupancy combination and incorporated into the “unknown” vulnerability curve for a location of this type.

Identifying bulk coding can be tricky for certain portfolios or lines of business. However, risk profiles and year-over-year comparisons can point out inconsistencies as well as large segments of the data with identical characteristics.

Learning Objective 33: Identify validation checks which would indicate potential erroneous data or poor data quality. (e.g. 100 story wood frame building)

There are a number of validation checks that can be performed on the data to verify clean, qualify data. Here is a listing of a few tests that could highlight data issues:

- Logic rules: Limits are greater than values or policy inception/expiration dates are outside of a reasonable date range.
- How do values compare to LOB/Construction/Occupancy/Square Footage: Is the value for a residential, wood frame, 1600 ft building or the cost of an automobile significantly higher than reasonable? $1M mobile home?
- How do number of stories compare to construction/occupancy: Do the number of stories make sense for the type of structure? E.g., 20 story wood frame or un-reinforced masonry construction.
- Invalid Construction/Occupancy combinations: These can be harder to detect but an example would be a wood frame construction with an industrial occupancy
- Year built: Outside normal parameters too old or in the future
- Outliers: Any data elements that are inconsistent or unexpected with others in that data field
- Completeness: High % unknown for any data element
- Duplication: Policy and location information are duplicated within the portfolio

The following article contains valuable information on data quality, including the table below on data quality rules:

7 Step Quality Process: https://www.moodysanalytics.com/risk-perspectives-magazine/managing-insurance-risk/insurance-regulatory-spotlight/data-quality-is-the-biggest-challenge

Type of rule	Description	Insurance example
Basic business rules	Basic business logic rules.	The date of a claim cannot be earlier than the date of policy inception.
Data-type constraints	Values in a particular column must be of a particular data type; for example, Boolean, numeric (integer or real), date, and code.	CIC/ISO Code in asset data for QRT templates.
Regulatory constraints	Data validations or rules laid down by regulators such as EIOPA, the Federal Financial Supervisory Authority (BaFin) in Germany, or the PRA.	The validation rules contained in the QRT templates specified by EIOPA.
Range constraints	Typically, numbers or dates should fall within a certain range. That is, they have minimum and/or maximum allowable values.	» Eligible ages for contracts – eligible age must be between, for example, 18-65. » Minimum premium amounts. » Reassurance limits.
Mandatory constraints	Certain columns cannot be empty (e.g., not null).	
Unique constraints	A field, or a combination of fields, must be unique across a dataset.	No two policyholders can have the same national insurance number.
Set-membership constraints	The values for a column come from a set of discrete values or codes.	For example, a person's gender may be Female, Male, or Unknown (not recorded).
Foreign-key constraints	This is the more general case of set membership. The set of values in a column is defined in a column of another table that contains unique values.	For example, in a "town" column, it must be aligned to a county from a "County Table."
Regular expression patterns	Occasionally, text fields will have to be validated this way.	For example, phone numbers may be required to have the (0044) 1234-5678 format.

Learning Objective 34: Explain the rationale behind assessing data quality and the effect it may have on the results. Provide peril specific examples.

As noted in Understanding Uncertainty in Catastrophe Modelling for Non-Catastrophe Modellers coding (Uncertainty in exposure data section): http://www.lmalloyds.com/AsiCommon/Controls/BSA/Downloader.aspx?iDocumentStorageKey=cc44f6be-b83f-4cf9-903c-e802c1f312a8&iFileTypeCode=PDF&iFileName=Understanding%20uncertainty%20in%20cat%20modelling%20for%20non-cat%20modellers , the accuracy of any assessment is reliant on the quality of the input data. Even where information is provided, there is the potential for it to be inaccurate. Considering the challenges of gathering and maintaining such a significant volume of data it is highly likely that any dataset contains inaccuracies.

Additionally, as noted above, systematic data issues or biases across a portfolio can result in losses being consistently under- or over-estimated. Identifying these data issues can help improve accuracy in modeling results. Essentially, data quality, accuracy and completeness can have substantial impacts on model results and is an important component of catastrophe modeling.

However, it is important to keep in mind that, for different classes of business, regions and perils, the quality of data available will vary and low data quality will have different impacts. For instance, the perils of terrorism and wildfire require accurate address information as the event footprint is both small and location specific. Slight shifts in location remove or add risks to high hazard areas.