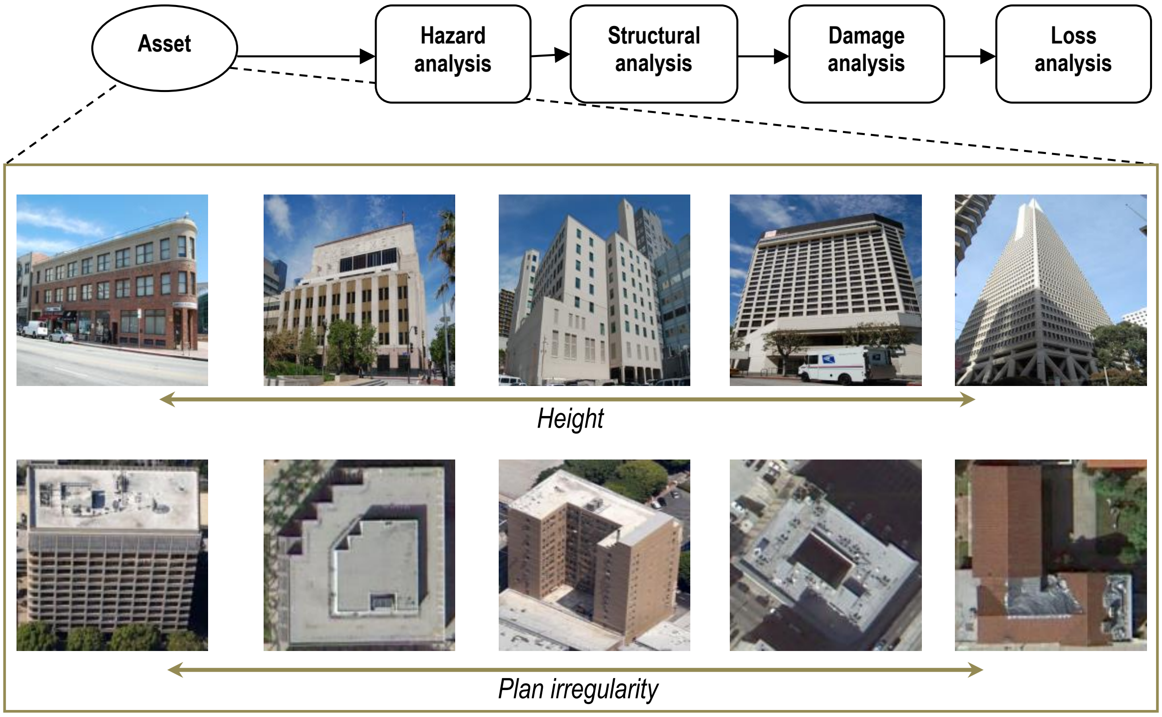
A Field Sampling Strategy for Estimating a Building Inventory

Report produced in the context of the

Global Earthquake Model Inventory Data Capture Tools

K. Porter[1]

[1] SPA Risk LLC, Denver CO USA



A Field Sampling Strategy for Estimating a Building Inventory

Global Earthquake Model Inventory Data Capture Tools

Version: 10

Responsible Coordinator: J. Bevington

Author: K. Porter

Date: September 3, 2013

The views and interpretations in this document are those of the individual author(s) and should not be attributed to GEM Foundation. With them also lies the responsibility for the scientific and technical data presented. The authors do not guarantee that the information in this report is completely accurate.

Citation: Porter, K., (2013) A Field Sampling Strategy for Estimating a Building Inventory, Inventory Data Capture Tools. Global Earthquake Model, http://nexus.globalquakemodel.org

Version history

Rev	Date	Comment
0	25 Jan 2012	Initial draft, following requirements specified in 17 Mar 2011 conversation with Huyck
1	28 Aug 2012	Simplify introduction, deleting references to other aspects of the sampling strategy such as pilot tests. Rewrite with less engineering jargon (“given, required, solution”). Minor wording changes. Move explanation of fancy math to appendices.
2	7 Sep 2012	Wording changes. Per call with Huyck & Z, made it clear that some attributes in the short list of building types are optional, e.g., era, roof, etc.
3	12 Oct 2012	Per WebEx w Bevington, Huyck, and Z, changed “block” to “cluster.” Added nonstratified (simpler Monte Carlo-type) sampling scheme for cases where zones have fairly uniform height. Added instructions on what to do if a selected building were missing or inaccessible. Added required metadata. Added placeholder for mapping scheme. Defined all terms. Re-read for newcomer. Made it clearer which tasks are performed by the inspector and which by SIDD. Added this revision history.
4	17 Oct 2012	Per WebEx w Bevington & Z today, change “district” to “zone” for consistency with GEM IDCT standard terminiology; add text near Table 5 and Table 6 suggesting that the data might be entered into this Word table, or directly into SIDD, or on a spreadsheet for which IDCT will create a template. In Sec 4.2 item 2, add UI suggestion to programmers about how to give user control over how much weight to give expert. Add Sec 4.4, Confirmation of Bayesian updating procedure. It is both a confirmation and illustration of the effect of sample size.
5	16 Apr 2013	Per Oct 2012 (?) telecon w Huyck, deleted placeholder for Sec 5 Mapping Scheme
6	24 Apr 2013	New intro table to summarize methods. Note on sample size in Sec 1.1. New Appendix A.3 on sample size to explain the math and inform tradeoffs between sample size and sensitivity to rare building types.
7	2 May 2013	Estimate productivity: ~1 zone per team-day. Clarify introduction. Pageinate.
8	8 May 2013	Move summary of sample size to main body (Sec 2) and clarify. Add new section on special buildings (Sec 8). Expand discussion of replacement buildings, in case preselected buildings are inaccessible.
9	2 Sep 2013	Formatted to approximately match GEM report template. Until published by GEM, the report made temporarily available at www.sparisk.com
10	3 Sep 2013	Formatted to match GEM report template. Submitted for publication at www.nexus.globalquakemodel.org

ABSTRACT

This memo offers procedures to estimate the area distribution of building types in a community through a combination of remote sensing, local expertise, and field observations. The resulting inventory can be used for probabilistic seismic risk analysis. The community is first divided into geographic regions (referred to here as zones) of relatively homogeneous use. These procedures show how to estimate the area distribution of building types for one zone. One estimates the total building area in each zone by remote-sensing procedures that are not discussed here, and then multiplies by the distribution estimated using the present procedures. The result is an estimate of the building area of each building type in each zone. One repeats the process for each zone in the community to create an estimate of the building area of each building type in the community. Four sampling strategies are offered for sampling a zone depending on whether (a) the zone has homogeneous building heights, (b) all important building features are visible, (c) expert advice is available on recognizing hidden features, and (d) a prior estimate from experts on building-type distribution is available. The strategies are:

1. Simple sampling without prior expert judgment. This procedure applies where the survey team can identify building type from visible features. It is useful for zones with homogeneous (fairly uniform) heights. That is, one cannot easily pick out clusters of buildings that are on average shorter, typical, or taller in terms of number of stories. It uses simple weighted averages to extrapolate field observations to the entire zone.
2. Stratified sampling without prior expert judgment. Like 1, except that the zone has a heterogeneous (not so uniform) mix of heights, e.g., a central business district with some generally low-, mid- and highrise blocks. It uses a procedure called moment matching to select the sample and extrapolate to the zone.
3. Use local expertise to infer types from visible features. Like 1 or 2, but a least one attribute used to define building type is not visible. Prior expert advice is needed to infer building type from visible features. Field data collectors then employ either simple or stratified sampling depending on height homogeneity. If the expert is willing to estimate the distribution of building types, this procedure shows how to do that.
4. Field survey to enhance prior expert opinion of type distribution. Like 3, except that it requires the expert to provide a prior distribution of building type by area, which is then combined with data from the field observations using Bayes' theorem.

It is estimated that one team of two people can survey enough buildings to represent the distribution of one zone in approximately one day. A separate procedure is offered for special buildings, such as buildings that serve some important function or are known to be particularly seismically vulnerable or particularly seismically resilient. Examples include hospitals, tall masonry towers, or base-isolated buildings. An appendix provides the mathematical basis of these procedures, including stratified sampling, Bayseian updating, and a discussion of the tradeoff between sample size and sensitivity to rare building types.

Keywords: inventory; sampling; survey; building type; distribution

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Most variables are defined near their first use. Other common abbreviations in this document are as follows.

CoV	Coefficient of variation (not the same as covariance)
GED4GEM	Global Exposure Database for the Global Earthqake Model
GEM	Global Earthquake Model
IDCT	Inventry Data Capture Tools
SIDD	Spatial Inventory and Damage Data tool

1. Introduction

Objectives. This memo offers procedures to estimate the area distribution of building types in a community through a combination of remote sensing, local expertise, and field observations, for later use in calculating seismic risk using the Global Earthquake Model (GEM). It is a product of GEM's Inventory Data Capture Tools (IDCT) project. The memo assumes that the community can be divided into geographic regions (referred to here as zones) of relatively homogeneous use. A zone might be the central business district, a manufacturing district, or a residential neighborhood. These procedures show how to estimate the area distribution of building types for one zone. One estimates the total building area in each zone (by procedures not discussed here) and multiplies by the estimated distribution to produce an estimate of the building area of each building type in each zone. One repeats the process for each zone in the community to create an estimate of the building area of each building type in the community.

Organization of the memo. Section 2 discusses sample size: how many buildings must be observed to achieve either of 2 sensitivity objectives. For convenience and simplicity this memo assumes a standard sample size, but the survey manager can vary the sample size considering the guidance in that section. Section 0 suggests that field data collection for one zone takes approximately 1 day for a team of 1 or 2 people. How one performs the sample and interprets the data depends on (a) whether building heights within a zone of homoegeous use are reletaively homogeneous, (b) whether the important building features are all visible, and (c) whether one can get an expert to estimate the building-type distribution in advance of the field data collection effort. See Table 1 for guidance on choosing among 4 sampling strategies depending on the answers to these questions. They are described in Sections 3, 4, 5, and 6 respectively. If it is known that there are unusual but important buildings that survey manager wants to be sure to include in the survey, see Section 7 for guidance. See Appendix A for background on some of the math employed here.

Table 1. Sampling strategies presented here

Homogeneous heights within a zone	All important building features are visible	Expert advice on recognizing hidden features	Expert advice on building-type distribution	Sampling strategy	See memo section
Yes	Yes	No	No	1	3
No	Yes	No	No	2	4
Either	No	Yes	No	3	5
Either	No	Yes	Yes	4	6

Summary of the sampling strategies. In more detail, sampling strategies 1 through 4 are as follows:

1. ***Simple sampling without prior expert judgment.*** This procedure applies where the survey team can identify building type from visible features. It is useful for zones with homogeneous (fairly uniform) heights. That is, one cannot easily pick out clusters of buildings that are on average shorter, typical, or taller in terms of number of stories. It uses simple weighted averages to extrapolate field observations to the entire zone.
2. ***Stratified sampling without prior expert judgment.*** Like 1, except that the zone has a heterogeneous (not so uniform) mix of heights, e.g., a central business district with some generally low-, mid- and highrise blocks. It uses a procedure called moment matching to select the sample and extrapolate to the zone.
3. ***Use local expertise to infer types from visible features.*** Like 1 or 2, but a least one attribute used to define building type is *not* visible. Prior expert advice is needed to infer building type from visible features. Field data collectors then employ either simple or stratified sampling depending on height homogeneity. If the expert is willing to estimate the distribution of building types, this procedure shows how to do that.
4. ***Field survey to enhance prior expert opinion of type distribution***. Like 3, except that it requires the expert to provide a prior distribution of building type by area, which is then combined with data from the field observations using Bayes' theorem.

2. Sample Size and Productivity

2.1 Required Sample Size

How large a sample size is needed? There is a tradeoff between small samples (which make the process efficient) and the desire to observe all common building types (which is desirable). Let us refer to a “survey manager,” a person who manages the survey and among other things chooses the survey objectives. Let us refer to a “sensitivity objective,” a quantitative objective about how confident the survey manager wishes to be to observe less-common building types. It is proposed here that the survey manager chooses one of two survey sensitivity objectives:

1. ***Observe a least one.*** In any homogenous zone, the survey shall observe with a specified confidence *f* (the survey manager chooses *f*, say 50%, 90%, or 95%) at least 1 sample of any building type *t* that occurs in some small fraction *p* of the population (the survey manager chooses *p*, say 10%, 5%, or 1%). This approach does not include an objective to quantify any variability of building features within type *t*.
2. ***Observe variability.*** In any homogenous zone, the survey shall observe with a specified confidence *f* (the survey manager chooses *f*, say 50%, 90%, or 95%) at least 3 samples of any building type *t* that in reality occurs in some small fraction *p* of the population (the survey manager chooses *p*, say 10%, 5%, or 1%). This approach allows one to quantify at least some variability of building features within type *t*. If more samples of each type are desired, see Appendix A.3.

If the sensitivity objective is to observe at least one, see Table 2 for the required sample size. If the sensitivity objective is to observe at least 3, see Table 3 for the required sample size. The tables follow automatically if buildings of each type *t* are randomly distributed through the zone. See Appendix A.3 for the math.

A few examples to make sure the meaning of Table 2 is clear. If the survey manager wants to have at least 2:1 odds ($f \geq 67\%$) of observing at least one building of a type with frequency $p = 0.05$, then we need to observe at least $N = 21$ of them (3 clusters of 7). See the cell with the row header labeled 7 per cluster ($N = 21$) under the column header labeled $p = 0.05$. The cell says $f = 66\%$, which is approximately 2:1 odds. With 3 clusters of 3, ($N = 9$), the survey is fairly likely (1 - 37% = 63% chance) to miss buildings that appear with frequency less than 1 in 20 ($p = 0.05$). With 3 clusters of 10 ($N = 30$), the survey has roughly even odds ($f = 45\%$ chance) of seeing at least one sample of any type that appears with frequency of at least 1 in 50 (p = 0.02).This memo recommends 3 clusters of 10 ($N = 30$) only because it seems convenient and easy to remember, but Table 2 can guide the survey manager who wants to trade off sample size and chance of missing a type.

An example from Table 3: With an objective of observing with at least even odds ($f \geq 50\%$) at least 3 samples of a building type that occurs in 1 of 10 buildings ($p = 0.1$), one needs to observe at least 9 buildings per cluster ($N = 27$), because this combination gives f = 52%.

Table 2. Likelihood of observing at least 1 specimen of type *t* among *N* samples

Sample size		Frequency with which type t occurs			
		1 in 100	1 in 50	1 in 20	1 in 10
Per cluster	*N*	*p* = 0.01	*p* = 0.02	*p* = 0.05	*p* = 0.1
1	3	3%	6%	14%	27%
2	6	6%	11%	26%	47%
3	9	9%	17%	37%	61%
4	12	11%	22%	46%	72%
5	15	14%	26%	54%	79%
6	18	17%	30%	60%	85%
7	21	19%	35%	66%	89%
8	24	21%	38%	71%	92%
9	27	24%	42%	75%	94%
10	30	26%	45%	79%	96%
20	60	84%	97%	100%	100%
30	90	100%	100%	100%	100%

Interpret 100% to mean “almost certain,” not “absolutely certain.”

Table 3. Likelihood of observing at least 3 specimens of type *t* among *N* samples

Sample size		Frequency with which type t occurs			
		1 in 100	1 in 50	1 in 20	1 in 10
Per cluster	*N*	*p* = 0.01	*p* = 0.02	*p* = 0.05	*p* = 0.1
1	3	0%	0%	0%	0%
2	6	0%	0%	0%	2%
3	9	0%	0%	1%	5%
4	12	0%	0%	2%	11%
5	15	0%	0%	4%	18%
6	18	0%	1%	6%	27%
7	21	0%	1%	8%	35%
8	24	0%	1%	12%	44%
9	27	0%	2%	15%	52%
10	30	0%	2%	19%	59%
20	60	2%	12%	58%	95%
30	90	6%	27%	83%	100%
40	120	12%	43%	94%	100%
50	150	19%	58%	98%	100%
60	180	27%	70%	99%	100%
70	210	35%	79%	100%	100%
80	240	43%	86%	100%	100%
90	270	51%	91%	100%	100%
100	300	58%	94%	100%	100%

Interpret 0% to mean “highly unlikely” and 100% to mean “almost certain,” not “absolutely certain.”

2.2 Productivity

How may buildings can a survey team (1 person or 2 people working together) examine in a day? A survey team can examine perhaps 5 buildings per hour when the buildings are on the same block, after the team has become adept at collecting and recording the data. Possibly the number is different, perhaps as few as 3 or 4, or as many as 6 or 7. Travel time between one cluster of buildings to another depends on the distance between the clusters. Let us assume clusters are separated by 15 minutes’ travel time.

It is later proposed that the survey team that wishes to characterize the distribution of building types within a zone of relatively homogeous use – say a central business disrict, a warehouse district, or a dense residential district – should examine 30 buildings. The 30 buildings comprise three clusters of 10 per cluster. The time required is the time to examine

30 buildings and to make 4 moves: one to the first cluster, one move between the 1st and 2nd cluster, one move between the 2nd and 3rd cluster, and one move to return from the field. The total time required is then:

$$T = 30 \text{ buildings}/(5 \text{ buildings per hour}) + 4 \text{ moves}\cdot 0.25 \text{ hr/move} = 7 \text{ hr}$$

Adding warm-up time and a meal break, this suggests that it takes approximately one team-day (8 hours) to perform the field data collection necessary to characterize the building-type distribution for one zone of relatively homoegeous use.

3. Sampling of a Height-Homogeneous Zone Without Prior Expert Judgment

In some cases the survey team possesses enough expertise to infer building types from building features that are visible from the street. This section specifies a procedure to make sample observations and infer the distribution of building types throughout a geographical area.

The following algorthim is applied to a single "zone," a geographic area supposed to have fairly homogeneous though not entire uniform construction. A zone can optionally be one or more grid cells as defined by GED4GEM. Unless it is perfectly uniform, it is supposed here that there are areas with larger buildings and areas with smaller buildings, but the building heights are so homogenous that one cannot pick out clusters of relatively short, typical, or tall buildings. The following procedure is repeated for each zone.

Sample size: There is a tradeoff between small samples (which make the process efficient) and the likelihood of missing a building type (which is undesirable). The minimum sample size depends on how confident one wants to be of observing rarer building types. This memo recommends sampling 3 clusters of 10 buildings in each zone for a total of 30 buildings per zone because it seems convenient and easy to remember. As it turns out, this sample size provides at least an even chance of seeing at least one sample of any type that appears with frequency of at least 1 in 50 ($p = 0.02$), and 95% confidence that one will observe at least one building of any type that represents at least 10% of the building stock by count. The analyst who wants to make a more deliberate tradeoff between sample size and sensitivity to rarer building types can see Appendix A.3.

Metadata. The following metadata items are needed.

- Field inspectors' names. One must be able to associate each building observation with the names of the inspectors who collected that observation.
- Field data tool name and version number
- Geographic boundaries of the zone. The inspectors do not need to know this.
- A unique identifier or numerical index for the zone. By "index" is meant here an integer label.
- A unique identifier or numerical index for each building observed in the zone.
- Date and local time of each building observation. Time can be the local time at which the inspectors arrived at the building. These metadata are obviously collected during the fieldwork, but are listed here for convenience.

Pre-fieldwork data. The following data items are needed before performing the field observation.

O_0, the geographic coordinates of cluster 0, selected spatially at random from the zone. The coordinates can be in terms of decimal degrees north latitude and east longitude, or can be a cross street, a milepost on a particular road, or any convenient coordinate system where the surveyor can know that he or she is standing within a few dozen meters of location O_0.

O_1, ditto, cluster 1

O_2, ditto, cluster 2

O_{00}, the spatial coordinates of the front door (or some accessible point at the perimeter of the building) of building 0 in cluster 0, selected at random from buildings near O_0. The coordinate system can again be in decimal degrees north latitude and east longitude or another system so long as the coordinates uniquely identify a building.

O_{01}, ditto, building 1 in cluster 0.

...

O_{09}, ditto, building 9 in cluster 0.

O_{10}, ditto, building 0 in cluster 1.

O_{11}, ditto, building 1 in cluster 1.

..

O_{19}, ditto, building 9 in cluster 1.

O_{20}, ditto, building 0 in cluster 2.

O_{21}, ditto, building 1 in cluster 2.

...

O_{29}, ditto, building 9 in cluster 2.

p = total plan area of buildings in the zone in square meters, determined from remote sensing or as defined by GED4GEM.

t = index to building types in the zone. For example, if there are 3 building types in the zone, they might have indices 0, 1, and 2. The indices must be consistent across all zones in the survey, so that t = 1 refers to the same building type in all zones.

The procedure will produce the following estimates:

1. Total building area by type and total for the zone
   E[A] = expected value of building area (square meters) in the zone
   E[A(t)] = expected value of building area (square meters) of type t in the zone
2. Total building count by type and zone total count
   E[N(t)] = expected value of the number of buildings of type t in the zone
   E[N] = expected value of the number of buildings in the zone

The procedure starts by picking three sample groups of buildings, one on each of the clusters at O_0, O_1, and O_2.

Pick the sample & record observations

1. For each cluster i and building j in cluster i, go to the building at location O_{ij}. Record the building's plan area, height (number of stories), and type t. Let us denote by P_{ij}, H_{ij}, and T_{ij} the plan area, height, and building type of building *j* in cluster *i*. Here, $i \in \{0, 1, 2\}$ and $j \in \{0, 1, 2, \dots 9\}$. The symbol $\in$ means "is one of" and the elipses "..." means "and all the next integers up to." The number of clusters and the number of buildings in the cluster could be themselves variables, but for simplicity we assume here that there are always 3 clusters and always 10 buildings in each cluster.
2. Inaccessible buildings: If the building at location O_{ij} is not accessible, select the nearest accessible building that was not already among the preselected buildings in the cluster. This is a "replacement building." Record its O_{ij}, overwriting the original location and noting in the comments that the building at the originally selected O_{ij} was not accessible. Record that building's plan area, height (number of stories), and type t instead of those of the inaccessible building. Again, select the *nearest accessible building*, without consideration of how complicated it is or any other feature.

Compile sample statistics

1. This and subsequent calculations are performed by SIDD, not the field inspector. Estimate building area of building j in cluster i. Let us denote this as A_{ij}. For simplicity, $A_{ij} = P_{ij} \text{ x } H_{ij}$. We recognize that a small fraction of buildings may be nonprismatic, e.g., a tall building on a lowrise platform with a much larger plan area. This memo ignores that case.
2. Calculate $N_0(t)$, number of buildings in cluster 0 of type t, ditto $N_1(t)$ in cluster 1, ditto $N_2(t)$ in cluster 2.
3. Calculate $P_0(t)$, total plan area of buildings in cluster 0 of type t; similarly $P_1(t)$ and $P_2(t)$. $P_0(t)$ is the simple sum of plan areas of those buildings in cluster 0 that are of type t.
4. Calculate $A_0(t)$, total building area of buildings in cluster 0 of type t; similarly $A_1(t)$ and $A_2(t)$. (Total building area is taken as plan area times height, summed over the observed buildings.). $A_0(t)$ is the simple sum of A_{0j} of buildings that are of type t.

Estimate zone-level quantities

5. Calculate weighted-average total number of buildings per cluster by type
$$E[N(t)/cluster] = 0.33*N_0(t) + 0.34*N_1(t) + 0.33*N_2(t)$$
6. Calculate weighted-average cluster-total plan area by type t, and total:
$$E[P(t)/cluster] = 0.33*P_0(t) + 0.34*P_1(t) + 0.33*P_2(t)$$
$$E[P/cluster] = Sum_t\{E[P(t)/cluster]\}$$
7. Calculate weighted average cluster-total building area by type t, and total:
$$E[A(t)/cluster] = 0.33*A_0(t) + 0.34*A_1(t) + 0.33*A_2(t)$$
$$E[A/cluster] = Sum_t\{E[A(t)/cluster]\}$$
8. Calculate estimated zone-level fraction of total area by type t
$$E[f(t)] = E[A(t)/cluster] / Sum_t\{E[A(t)/cluster]\}$$
9. Calculate weighted-average total building area per building by type
$$E[G(t)] = E[A(t)/cluster] / E[N(t)/cluster]$$
10. Calculate sample-average height of buildings, weighted by building area
$$E[H] = E[A/cluster] / E[P/cluster]$$
11. Estimate total building area in the zone and total by type
$$\boxed{\begin{array}{c} E[A] = p * E[H] \\ E[A(t)] = E[A] * E[f(t)] \end{array}}$$
12. Estimate total number of buildings zone-wide by type

$$E[N(t)] = \mathrm{round}(E[A(t)]/E[G(t)])$$

$$E[N] = \mathrm{Sum}_t\{E[N(t)]\}$$

4. Sampling of a Height-Heterogeneous Zone Without Prior Expert Judgment

This method is like method 1, except that the zone is heterogeneous enough that one can pick out from remote-sensing imagery clusters of short, medium, and tall buildings. In this case, for purposes of improving the sample accuracy, we stratify the zone by average building height within small spatial clusters of buildings. This is called a stratified sample, meaning that the sample is stratified--separated into contiguous ranges—by building height. For example, “short” might be buildings of 1 story, “medium” might be buildings of 2-4 stories, and “tall” might be buildings of 5 or more stories. The strata do not need to be consistent across zones or across different surveys. The following procedure is repeated for each zone.

4.1 Sampling Procedure

The procedure requires that the following information be known:

H_0 = lower bound (10th percentile) of the average number of stories of a cluster of buildings in the zone. That is, approximately 10% of total building area is in a building of h_l stories or less. Determined by visual inspection of imagery. By “approximately” is meant that the 10th, 50th, and 90th percentiles can be judged by eye. One does not need to rigorously calculate and invert a cumulative distribution function. That sort of precision, while possibly valuable, may only occasionaly be practical and is not expected here.
H_1 = median (50th percentile) stories of a cluster in the zone. From visual inspection of imagery.
H_2 = upper bound (90th percentile) stories of a cluster in the zone. From visual inspection of imagery.
p = total plan area of buildings in the zone, determined from remote sensing or as defined by GED4GEM.
t = index to building types in the zone.

The procedure will produce the following estimates:

1. <u>Total building area by type and total for the zone</u>
   E[A] = expected value of building area (square meters) in the zone
   E[A(t)] = expected value of building area (square meters) of type t in the zone
2. <u>Total building count by type and zone total count</u>
   E[N(t)] = expected value of the number of buildings of type t in the zone
   E[N] = expected value of the number of buildings in the zone

Metadata. The following metadata items are needed.

- Field inspectors’ names. One must be able to associate each building observation with the names of the inspectors who collected that observation.
- Field data tool name and version number
- Date and local time of each building observation. Time can be the local time at which the inspectors arrived at the building. These metadata are obviously collected during the fieldwork, but are listed here for convenience.

Pre-fieldwork data. The following data items are needed before performing the field observation.
O_0, the geographic coordinates of cluster 0, a cluster of buildings with height H_0, selected spatially at random from among all clusters with average number of stories H_0.
O_1, ditto, cluster 1. That is, O1 is the location of a cluster of buildings with average height H_1, selected spatially at random from all clusters with average number of stories H_1.
O_2, ditto, cluster 2.
O_{00}, the spatial coordinates of the front door (or some accessible point at the perimeter of the building) of building 0 in cluster 0, selected at random from buildings near O_0.
O_{01}, ditto, building 1 in cluster 0
...
O_{09}, ditto, building 9 in cluster 0.
O_{10}, ditto, building 0 in cluster 1.
O_{11}, ditto, building 1 in cluster 1.
..

O_{19}, ditto, building 9 in cluster 1.
O_{20}, ditto, building 0 in cluster 2.
O_{21}, ditto, building 1 in cluster 2
...
O_{29}, ditto, building 9 in cluster 2.
p = total plan area of buildings in the zone, determined from remote sensing or as defined by GED4GEM.
t = index to building types in the zone.

Pick the sample & record observations

1. For each cluster i and building j in cluster i, go to the building at location O_{ij}. Record the building's plan area, height (number of stories), and type t. Let us denote by P_{ij}, H_{ij}, and T_{ij} the plan area, height, and building type of building *j* in cluster *i*. Here, $i \in \{0, 1, 2\}$ and $j \in \{0, 1, 2, ... 9\}$. The symbol $\in$ means "is one of" and the elipses "..." means "and all the next integers up to." The number of clusters and the number of buildings in the cluster could be themselves variables, but for simplicity we assume here that there are always 3 clusters and always 10 buildings in each cluster.
2. Inaccessible buildings: If the building at location O_{ij} is not accessible, select the nearest accessible building that was not already among the preselected buildings in the cluster. This is a "replacement building." Record its O_{ij}, overwriting the original location and noting in the comments that the building at the originally selected O_{ij} was not accessible. Record that building's plan area, height (number of stories), and type *t* instead of those of the inaccessible building. Again, select the *nearest accessible building*, without consideration of how complicated it is or any other feature.

Compile sample statistics

1. This and subsequent calculations are performed by SIDD, not the field inspector. Estimate building area of building j in cluster i. Let us denote this as A_{ij}. For simplicity, $A_{ij} = P_{ij} \times H_{ij}$. We recognize that a small fraction of buildings may be nonprismatic, e.g., a tall building on a lowrise platform with a much larger plan area. This memo ignores that case.
2. Calculate $N_0(t)$, number of buildings in cluster 0 of type t, ditto $N_1(t)$ in cluster 1, ditto $N_2(t)$ in cluster 2.
3. Calculate $P_0(t)$, total plan area of buildings in cluster 0 of type t; similarly $P_1(t)$ and $P_2(t)$. $P_0(t)$ is the simple sum of plan areas of those buildings in cluster 0 that are of type t.
4. Calculate $A_0(t)$, total building area of buildings in cluster 0 of type t; similarly $A_1(t)$ and $A_2(t)$. (Total building area is taken as plan area times height, summed over the observed buildings.). $A_0(t)$ is the simple sum of A_{0j} of buildings that are of type t.

Estimate zone-level quantities

1. Calculate weighted-average total number of buildings per cluster by type. ***Note to programmers***: we are creating a weighted average, and unlike the previous method, the weights in this procedure are unequal.
$$E[N(t)/cluster] = 0.3*N_0(t) + 0.4*N_1(t) + 0.3*N_2(t)$$
2. Calculate weighted-average cluster-total plan area by type t, and total:
$$E[P(t)/cluster] = 0.3*P_0(t) + 0.4*P_1(t) + 0.3*P_2(t)$$
$$E[P/cluster] = Sum_t\{E[P(t)/cluster]\}$$
3. Calculate weighted average cluster-total building area by type t, and total:
$$E[A(t)/cluster] = 0.3*A_0(t) + 0.4*A_1(t) + 0.3*A_2(t)$$
$$E[A/cluster] = Sum_t\{E[A(t)/cluster]\}$$
4. Calculate estimated zone-level fraction of total area by type t
$$E[f(t)] = E[A(t)/cluster] / Sum_t\{E[A(t)/cluster]\}$$
5. Calculate weighted-average total building area per building by type
$$E[G(t)] = E[A(t)/cluster] / E[N(t)/cluster]$$
6. Calculate sample-average height of buildings, weighted by building area
$$E[H] = E[A/cluster] / E[P/cluster]$$
7. Estimate total building area in the zone and total by type
$$E[A] = p * E[H]$$
$$E[A(t)] = E[A] * E[f(t)]$$
8. Estimate total number of buildings zone-wide by type
$$E[N(t)] = round(E[A(t)]/E[G(t)])$$
$$E[N] = Sum_t\{E[N(t)]\}$$

4.2 Sample Calculation 1, Imaginary Data

Imagine a zone with 10 clusters of 10 buildings each. Inspector will sample 3 clusters of 10 buildings. Each building may fall into 1 of 3 types t ∈ {X, Y, Z}. These types have plan area P and building height H as follows:

Table 4. Characteristics of building types. E[H] and CoV[H] ferfer to the expected value and coefficient of variation of building height of one building, respectively. Individual buildings' hieght and plan area are taken in this illustration as Gaussian and indepenently distributed, except that H is bounded below by 1 (e.g., no 0.7-story buildings). Frational number of stories (e.g., 1.3) should be interpreted as meaning that an upper floor has 30% the plan area of a lower floor.

	X	Y	Z
E[H]	1.2	1.5	2
E[P]	1000	1500	2000

	X	Y	Z
CoV[H]	0.1	0.1	0.1
CoV[P]	0.1	0.1	0.1

A sample calculation is shown next.

Table 5. Building inventory. Green highlighted cells are the observed sample and values that are calculated from it. Non-highlighted cells are not observable to the inventory collector. Yellow are the actual cumulative distribution functions of cluster-average building height, showing that clusters 2, 6, and 9 are the appropriate targets for clusters l, m, and u. Total plan area (row labeled "Total," column labeled "Total P") is observed from remote sensing

	N			P			A			Totals				
Cluster	X	Y	Z	X	Y	Z	X	Y	Z	N	P	A	h	FH(h)
1	8	1	1	7071	1524	1179	7499	1604	1179	10	9774	10283	1.05	0.046
2	7	2	1	5174	4542	1281	5174	7675	3661	10	10996	16510	1.50	0.120
3	7	1	2	8872	1748	4638	9089	2928	7989	10	15259	20006	1.31	0.210
4	5	3	2	6246	3177	3455	9230	6091	7222	10	12879	22543	1.75	0.311
5	4	4	2	923	6655	800	1868	12249	2311	10	8378	16427	1.96	0.385
6	4	3	3	2606	4156	5102	3449	6491	11490	10	11864	21431	1.81	0.481
7	3	4	3	2296	5201	10291	2691	8188	16272	10	17788	27151	1.53	0.603
8	2	4	4	3589	6059	9864	4840	6683	21606	10	19513	33128	1.70	0.752
9	1	5	4	1049	13088	7277	1388	23297	10910	10	21413	35596	1.66	0.911
10	0	5	5	0	8627	5569	0	9091	10680	10	14196	19770	1.39	1.000
Total	41	32	27	37826	54777	49456	45228	84297	93320	100	142060	222845	1.57	

These values produce the following results.

	X	Y	Z		Total
E[N(t)]/cluster	4	3.3	2.7		
E[P(t)]/cluster	2909	6951	4608	E[P]/cluster	14,468
E[A(t)]/cluster	3348	11888	8967	E[A]/cluster	24,204
E[f(t)]	0.14	0.49	0.37		
E[G(t)]/building	837	3,602	3,321		
				E[H]	1.67

True totals, zone-wide

	X	Y	Z		All
P(t)	37826	54777	49456	P	142,060
A(t)	45228	84297	93320		222,845
N(t)	41	32	27		100
f(t)	0.20	0.38	0.42		
H					1.57

Estimated total building area zone-wide

E[A] = p*E[H]	237,650
True A	222,845
Ratio estimate/true	1.07

Estimated total building area zone-wide by type

	X	Y	Z
E[A(t)] = E[A]*E[f(t)]	32,876	116,726	88,048
True A(t)	45,228	84,297	93,320
Ratio estimate/true	0.73	1.38	0.94

Estimated total number of buildings zone-wide by type

	X	Y	Z	Total
E[N(t)] = E[A(t)]/E[G(t)]	39	32	27	98
N(t)	41	32	27	100
Ratio estimate/true	0.96	1.01	0.98	0.98

Check of bias and variability in error

A small Monte Carlo simulation was carried out on this sample calculation, varying H and P, though not N. Purpose was solely to check whether the method produces bias. It doesn't; mean ratio of estimated quantity to actual quantity was generally 0.97 to 1.03, depending on type or total. The CoVs of the ratio for total estimated building area by type and total estimated count of buildings were approximately 0.2 and 0.1, respetively, and error in total zone building area and count both have CoVs of approximately 0.1. This test can be repeated with case-stury data to check bias accuracy with real data.

Table 6. Ratio of estimated quantity to actual quantity in 20 simulations of the sample calculation.

Sim	Type area				Number of buildings			
	X	Y	Z	Total	X	Y	Z	Total
1	0.90	1.05	1.37	1.16	0.90	0.95	0.93	0.93
2	0.92	1.03	1.02	1.00	0.89	0.94	0.91	0.91
3	1.14	1.07	0.73	0.97	1.09	1.15	1.12	1.12
4	1.20	0.76	1.31	1.10	0.77	0.81	0.79	0.79
5	1.65	0.95	0.63	0.95	1.14	1.21	1.17	1.17
6	0.93	0.89	0.92	0.91	0.94	1.00	0.97	0.97
7	1.11	0.80	1.09	1.00	1.00	1.06	1.03	1.03
8	0.95	1.15	1.10	1.08	0.91	0.96	0.93	0.93
9	0.68	1.08	1.19	1.04	0.93	0.98	0.95	0.95
10	1.33	1.31	1.17	1.24	0.95	1.01	0.98	0.98
11	1.01	1.01	1.29	1.14	0.91	0.97	0.94	0.94
12	1.02	1.23	0.71	0.95	1.06	1.12	1.08	1.08
13	0.83	0.93	1.10	0.97	0.94	0.99	0.96	0.96
14	0.85	0.88	1.15	0.99	0.93	0.98	0.95	0.95
15	0.84	1.00	0.75	0.85	0.97	1.02	0.99	0.99
16	0.80	1.34	1.24	1.20	0.92	0.97	0.94	0.94
17	0.77	0.73	1.11	0.93	0.93	0.98	0.95	0.95
18	1.25	1.16	0.96	1.07	1.03	1.09	1.05	1.05
19	0.85	1.08	1.15	1.06	0.77	0.81	0.79	0.79
20	0.82	1.24	1.06	1.06	1.13	1.20	1.16	1.16
Mean	0.99	1.03	1.05	1.03	0.96	1.01	0.98	0.98
CoV	0.23	0.17	0.20	0.10	0.10	0.10	0.10	0.10

Check of improvement in accuracy versus simpler Monte Carlo simulation

The same check was performed again without a stratified sample, i.e., with 3 clusters selected at random. Results are shown in Table 7. The table shows that with a simple Monte Carlo simulation (MCS), area and building-count estimates are still unbiased, but error CoVs are higher under MCS by about 1.5x versus a stratified sample. This demonstrates that stratified sampling is better than random sampling, at least with this toy case. It generally should, as long as the thing we stratify on – such as building height -- correlates with building type, which it seems as if it often will do. The degree of improvement from a stratified sample versus a random sample (here a 1/3rd reduction in CoV) will vary, probably depending on zone size, heterogeneity, and other factors. Note also that the stratification costs only a few minutes examining aerial or spatial imagery to spot short, medium, and tall clusters within the zone, and any incremental travel time to examine those particular places rather than other, possibly more-convenient, ones.

Table 7. Error results using simple Monte Carlo simulation (MCS). Table contents mean the same things as in Table 3. Note the higher CoVs, which show that MCS in this example produces a more-uncertain estimate of areas and building counts, though still unbiased in this case.

Sample	Type area				Number of buildings			
	X	Y	Z	Total	X	Y	Z	Total
1	0.59	1.40	1.09	1.08	0.60	1.00	0.97	0.83
2	0.70	0.94	0.93	0.88	0.63	1.03	0.92	0.84
3	1.21	0.74	1.17	1.02	0.89	0.98	1.05	0.96
4	0.89	1.05	0.83	0.90	1.23	1.12	1.33	1.22
5	0.57	1.14	1.16	1.05	0.54	0.98	0.97	0.80
6	1.52	0.51	0.98	0.94	1.97	0.85	0.84	1.29
7	1.00	0.94	0.99	0.97	1.06	0.90	0.94	0.98
8	0.58	1.38	0.94	0.98	0.63	1.03	0.91	0.84
9	0.89	0.95	1.19	1.04	0.84	1.03	0.97	0.94
10	1.37	0.62	1.25	1.08	1.20	0.90	1.20	1.10
11	0.90	1.17	1.23	1.15	0.81	0.98	1.16	0.96
12	1.65	1.07	0.73	1.06	1.31	0.80	0.71	0.98
13	1.38	0.82	0.91	0.96	1.28	0.89	1.06	1.09
14	0.75	0.90	1.19	1.01	0.92	1.12	0.96	1.00
15	1.27	0.91	0.87	0.98	1.61	0.86	1.02	1.20
16	0.96	1.32	0.82	1.03	0.82	1.01	0.95	0.92
17	0.90	1.33	0.64	0.86	0.82	1.34	1.19	1.09
18	0.91	1.23	0.73	0.96	0.80	1.19	1.03	0.99
19	0.86	0.98	1.13	1.03	0.87	0.79	0.94	0.86
20	1.38	0.82	0.79	0.88	1.05	0.73	0.87	0.90
Mean	**1.01**	**1.01**	**0.98**	**0.99**	**0.99**	**0.98**	**1.00**	**0.99**
CoV	**0.32**	**0.25**	**0.19**	**0.08**	**0.36**	**0.15**	**0.14**	**0.14**

5. Use Local Expertise to Infer Types from Visible Features

Some attributes of the GEM basic taxonomy are not visible, and in many cases cannot be confidently guessed from visible features by structural engineers who are unfamilar with local construction. By "attribute" is meant a feature that helps to distinguish one building type from another, such as the number of stories or the lateral load resisting system. (See GEM taxonomy documents for detail). Number of stories generally is observable, lateral load resisting system often isn't, even to experienced structural engineers, because for example it is concealed by architectural finishes. This procedure is designed to elicit local builder or building-department expertise on how to infer building type from visible features.

Invitation. The analyst must identify a building official capable of providing local expertise on the history of construction in the target community. We provide no guidance on how to do that, other than to contact the local building department and request contact with experienced building officials. The following text can be inserted in a memo to the local official, explaining what the meeting is intended to achieve.

Purpose of the discussion. A seismic risk analysis will be performed on buildings in your community, and we would like to solicit your expertise. In doing a seismic risk analysis of the kind we have in mind, we will create a mathematical model of the kinds of buildings exposed to earthquakes. You have been identified as an expert on construction practices and history in your community, so we would like to discuss local building types with you in a brief meeting. In the meeting, we will begin by explaining how we want to categorize the buildings. Some features that matter to the seismic performance of buildings can be hard to see, such as the lateral load resisting system. We want to find out from you whether there are visible clues to the hard-to-see features. It is common for a neighborhood or a city to have only a limited number of very common building types, perhaps 10-20. We would like to know whether you agree that that is true in your community, and what you think those common types are. Finally, we will ask you to help us create a table to relate combinations of visible features to the most common types. Time permitting, we might repeat the exercise by dividing the community up by zones.

Following are steps for the meeting with the local official.

Summarize GEM basic taxonomy. The analysis will use software developed by the Global Earthquake Model (www.globalquakemodel.org). The software requires information about buildings in terms of up to 8 attributes: structural material, lateral load-resisting system, roof material, floor material, height, era of construction, degree of vertical or plan irregularities, and occupancy. Show the local official the tables of each attribute, taken from the GEM basic taxonomy.

Get the short list. Ask the local official to try to identify a short list of perhaps 10 to 20 most-common combinations of the attributes of the GEM basic taxonomy. One way to document these combinations is to list them in Table 8 below. It may be that IDCT offers a user interface in SIDD so that the local official and interviewer enter the short list directly into SIDD. Or it may be that IDCT prepares a spreadsheet template that can be shared with the local official, and SIDD has the capability to import the filled-in spreadsheet.

In any case, in Table 8, "Fraction" refers to the fraction of all buildings (by count) that the expert thinks the given combination represents. It is only needed if (a) you want to skip the field survey entirely, or (b) you want to do Bayesian updating, discussed Section 6. If the fractions in this column do not sum to 100%, divide each one by the sum of them all, and replace the fractions with these normalized values.

Comunity:_______________________________________ Expert: ________________________Date:___________

Table 8. Short list of building types

	Occupancy(a)	Material	LLRS	Height	Era(b)	Irregularity(b)	Roof(b)	Floor(b)	Fraction
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
11									
12									
13									
14									
15									
16									
17									
18									
19									
20									

(a) Will be fairly uniform for a zone
(b) Complete if possible

Key visible attributes. Some attributes are hard to see: structural material, lateral load resisting system, roof, floor. What are 1, 2 or 3 key visible attributes that tell the most about the hidden ones? For example, if you could see height, occupancy, and the front wall finish, would those give you a good idea of the structural system, laterial load resisting system, floor and roof type? Let "Visible attribute 1" denote the 1st key visible attribute, "Visible attribute 2" denote the 2nd, etc.

The key attributes do not have to all be the names of attributes in the GEM Basic Building Taxonomy, Appendix A, Table G1, column labeled "Attribute." If they are, then you can skip this next step, since the entries in Table 5 tell you everything you need to know to estimate the probability distribution of the building type for any given combination of visible attributes. Reason is that the probability that any individual building is of type t is its probabilty in Table 5, divided by the sum of the probabilities of all entries in Table 5 that have the same key attribute values.

Visible attribute 1: ________________________________
Its most likely values
1a: ____
1b: ____
1c: ____
(Add rows if necessary, but try to keep the list as short as possible. If there are only 1 or 2 likely values, that is easier to handle than 3 or 4.)

Visible attribute 2: ________________________________
Its most likely values
2a: ____
2b: ____
2c: ____

Visible attribute 3: ________________________________

Its most likely values
3a: ____
3b: ____
3c: ____

Choose whether you want to do a deterministic or a probabilistic mapping. A deterministic mapping is where the expert is telling you just the one most-likely building type given each set of visible features. A probabilistic mapping is where the expert is telling you that two or more building types are possible for at east one set of visible features, and the expert is telling you the chance of each possible type.

Make a table (like Table 9, or in SIDD, or in a spreadsheet whose template IDCT will prepare) with one row for each unique and reasonable combination of the values of the 3 visible attributes. For each row, ask the expert to name the 1, 2, or 3 most likely types from Table 5, with the expert's guess of the likelihood of each one. That is, the entries in the column labeled "Vis 1" should be possible values of whatever attribute is named just above under "Visible attribute 1," i.e., it should contain values listed in 1a, 1b, etc. The entries in the column labeled "Best 1" should be the row number from Table 5, containing the short list of building types. The entry in the colum labeled "Prob 1" should be the expert's judgment of the likelihood that, if he or she saw the attributes listed under Vis 1 through Vis 3, that the building type would turn out to be the type identified under "Best 1." It is okay to just list 1 or 2 most likely types in a row.

Table 9. Mapping table from visible attributes to structure types

Row	Vis 1	Vis 2	Vis 3	Best 1	Fill these columns only for a probabilistic mapping				
					Prob 1	Best 2	Prob 2	Best 3	Prob 3
1									
2									
3									
4									
5									
6									
7									
8									
9									
10									
Etc.									

Do the field data collection based on the visible attributes of Table 9. That is, observe the 3 visible attributes mentioned above. For each building, either call it the type labeled under "Best 1," or it IDCT tools allow a probabilistic mapping, call it Best 1 with probability Prob 1, Best 2 with probability Prob 2, etc.

6. Sampling with Prior Expert Opinion of Type Distribution

This procedure shows how to get expert judgment of the building-type distribution and then enhance that judgment with a field survey. Sample calculations are also presented. The procedure employs Bayesian updating. For a brief explanation of Bayesian updating, see Appendix A, section A.2.

6.1 Overview of the Updating Procedure

Start with the following expert-judgment information (called the prior in Bayesian updating). Let

μ = expert's best estimate of the fraction of buildings (by count) within the community that are of type t
k = the number of buildings of type t observed in the field
n = the total number of buildings (all types) observed in the field

The fraction of buildings that are of type t are assumed to be uncertain. Let p denote that uncertain fraction. A convenient form for the probability distribution of p is the Beta distribution, bounded below and above by 0 and 1 respectively. In addition to the bounds, the parameters of the Beta distribution are α and β. The initial estimate for α and β are

$$\alpha = \nu \cdot \mu$$
$$\beta = (1-\mu)\cdot \nu \qquad (1)$$

The parameter v has to do with how strongly we believe the expert. A higher value equates with greater belief. A reasonable range might be v = 10 for a fairly inexperienced expert or someone who is only familiar with some of the building types in the community, v = 100 for one with a lot of experience and a good sense of probability. A Absent other information, take v = 50; this will make the degree of belief in expert's judgment roughly equal to observing 50 buildings in the community.

After making the field observations and finding that k out of n observations are of type t, α and β are updated to:

$$\alpha' = \alpha + k$$
$$\beta' = \beta + n - k \qquad (2)$$

After making the observations, the new best estimate of the fraction of buildings in the zone that are of type t is given by

$$\mu' = \frac{\alpha'}{(\alpha' + \beta')} \qquad (3)$$

6.2 Steps to Carry Out the Updating Procedure

Here is the algorithm for doing Bayesian updating.

1. Get the expert's short list (Table 5).
2. Assume a value for v, such as 50. This can be done by SIDD. ***UI suggestion for SIDD programmers***: maybe leave the choice to the user, say with a tab labeled "Bayesian updating." It has a table with a row for each type, and columns labeled "expert's judgment" "observed distribution" and "updated distribution." There could be a slider labeled "How much weight to give to expert versus observations." Label the left end of the slider "observations" and "expert" on the right end. Equate the left end with say *v* = 10, the right end with *v* = 100. Adjust the quantities in the updated distribution as the slider is moved. There could be a button to paste the results back into the mapping scheme tab.
3. Make the field observations. Let k_t denote the number of observations that are of type t. Let n denote the total number of observations made. Note that the the values *k* must sum to *n*.
4. For each building type, calculate α and β using Equation (1). This and all subsequent calculations are performed by SIDD.
5. For each building type, calculate α' and β' using Equation (2)
6. For each building type, calculate μ' using Equation (3). Use this value in loss estimation. This is the updated fraction of buildings by count that are represented by type t.

7. If you make additional observations in the same community, repeat steps 3 through 6, updating α' to α'' and β' to β'', and μ' to μ''. You do not need v. You only needed it to get an initial α and β.

6.3 Sample Calculation

Suppose the expert says there are 4 dominant building types, which are denoted below by t = 1, 2, 3, and 4. The expert says they are present in the community in the proportions 0.6, 0.2, 0.1, and 0.1 respectively. You then observe 30 buildings, of which 15, 9, and 3 are of types 1, 2, and 3. You observe no type-4 buildings, but you observe 3 that are of a type the expert did not mention; let us call that type 5.

For type 1,
$\alpha = \mu \cdot v = 50{\cdot}0.6 = 30$
$\beta = (1 - \mu) \cdot v = (1 - 0.6) \cdot 50 = 20$
$\alpha' = \alpha + k = 30 + 15 = 45$
$\beta' = \beta + n - k = 20 + 30 - 15 = 35$
$\mu' = \alpha'/(\alpha' + \beta') = 45/(45 + 35) = 0.56$

For the other types, see Table 10. Notice that μ moved from the expert's judgment of μ part way toward k/n. The lower v (your confidence in the expert) or the larger n (the amount of your evidence), the more μ would have movde toward k/n.

Table 10. Sample calculations for Bayesian updating

Type	μ	v	α	β	k	α'	β'	μ'
1	0.6	50	30	20	15	45	35	0.56
2	0.2	50	10	40	9	19	61	0.24
3	0.1	50	5	45	3	8	72	0.10
4	0.1	50	5	45	0	5	75	0.06
5	0	50	0	50	3	3	77	0.04
				n =	30		Check sum:	1.00

6.4 Confirmation of Bayesian Updating Procedure

Let us confirm that the Bayesian updating method works, and illustrate the effect of sample size.

Ler us begin with the same assumptions as in the sample calculation. Suppose in addition that the true distribution of types is as shown in Table 11, in the column labeled μ''. Let us show that with more observations, μ' converges to μ''. First some background math. We are making observations of a population (here, all the buildings in a zone) with some trials (here, observations of sample buildings within the zone). In the zone at any given time, there is some real distribution of buildings by building type within the zone. With an unbiased sample, the chance is μ_t'' that any given trial (any given building observed) results in a building of type *t* being observed. If there is more than possible outcome for any given observation (more than one building type in zone), and we observe fewer samples than the entire population, then any given sample will have an uncertain (a random) fraction of buildings in it of type *t*. The more samples one observes (the more buildings the field investigators observe), the closer the fraction of buildings observed of type t should move away from *μ* and toward μ_t''. The distribution of the total fraction observed of each type *t* as a multinomial distribution (http://en.wikipedia.org/wiki/Multinomial_distribution).

Without showing all the math here, I tested the procedure assuming that the real distribution of types was {0.80, 0.10, 0.05, 0.025, 0.025}.I then tested to see how close μ' (the updated distribution) moved from μ (the expert's udgment) to μ'' (the true distribution). I tested assuming that 10 buildings were observed, then 100, then 1000. The results should converge on μ'' with more observations, and that is what happened. See Table 11. Look at the columns labeled *E*[*μ*']. They column contains the average value of μ' after the specified number of observations *n*. The quantity *μ*' is random because the field investigators can look at a different sample of 10, 100, or 1000 buildings each time. Because *μ*' is random, it has an expected value *E*[*μ*'], standard deviation, and higher moments. As *n* increases, *E*[*μ*'] should approach *μ*''. That is what happens. These results take *v* = 50.

Table 11. Test of Bayesian updating procedures

Type t	Expert µ	True µ''	After Bayesian updating *E*[*µ*']		
			n = 10	*n* = 100	*n* = 1000
1	0.6	0.80	0.633	0.733	0.790
2	0.2	0.10	0.183	0.133	0.105
3	0.1	0.05	0.092	0.067	0.052
4	0.1	0.025	0.088	0.050	0.029
5	0.0	0.025	0.004	0.016	0.024

7. Special Building Types

It may be that there are special building types known to occur rarely in the community, but that are important. Perhaps they serve some important function or are known to be particularly seismically vulnerable or particularly seismically resilient. Examples include hospitals, tall masonry towers, or base-isolated buildings. In such a case, a person familiar with these buildings must identify them in advance, or at least identify the geographic area where they occur. If the buildings can be individually identified and there is a large number of them (too many to inventory each special building), or if one can delineate a geographic area that primarily includes these special buildings, exclude them from the surrounding zones and carry out the sampling of the surrounding zones as described earlier. Then sample the zone that includes these buildings as a separate zone. If there are few enough special buildings that they can be individually examined, exclude them from the surrounding zones and carry out the sampling of the surrounding zones as described earlier. Then observe each special building and treat each as a separate zone.

Appendix A. Mathematical Basis for these Procedures

A.1 Stratified Sampling

It is sometimes necessary in catstrophe risk modeling to perform difficult numerical integration, such as integrating some uncertain variable like number of stories over a large geographic region during the estimation total building area in the region. It can be a challenge to do that when automated methods do not exist or it is otherwise impractical to make all the necessary observations. For example it can be impractical to observe the typology and number of stories for each building in a zone. Instead of a complete integration then, we use a few discrete sample values of the uncertain variable, and associate each sample with a weight that approximates the probability distribution associated of the distributed uncertain variable.

Here, IDCT can estimate total plan area *p* of buildings in the zone, and needs to integrate over the distributed variable of building height to get uncertain total building area, A. Instead of multiplying each building's building plan area by its particular number of stories, we estimate the mean and standard deviation of height from these samples. Furthermore, we use the sample distribution of a second variable, building type, and integrate over building height, to estimate total building area by building type.

Moment matching is a generalization of Gaussian quadrature. This memo uses 3-point quadrature, and for simplicity only attempts to match the 1st 2 moments of a single scalar distribution, that of building height. That is, we use samples of building typology in 3 small areas within a zone: one area with low height, one with medium height, one with tall height. Also for simplicity, sample points and weights are chosen so that sample points have an easily understood cumulative probability. These are the 10th, 50th, and 90th percentiles, which although they have a clear quantitative meaning, also have an intuitive and easily communicated meaning, i.e., "lower bound," "typical value" and "upper bound." For more information on the latter, see http://en.wikipedia.org/wiki/Gaussian_quadrature.

A.2 Bayesian Updating

We want to know the probability that an arbitrary building in a zone or comminuty belongs to a particular type t. Sometimes we can get local experts to estimate that probability, and then use field sampling to update that expertise.

We use Bayesian updating to combine expert judgment of the distribution of buildings types with field observations. See http://en.wikipedia.org/wiki/Bayesian_updating for details. Briefly, some parameter of interest is taken as uncertain, having an initial probability distribution called a prior distribution. One then makes some observations that provide evidence about the parameter. An equation referred to as Bayes' theorem expresses how the probability distribution should rationally change to account for evidence. The changed probability distribution is called the posterior distribution. Bayesian updating is the only way to rigorously combine initial judgment about the probability distribution of some uncertainty quantity (like the distribution of building types) with observations, and produce an updated probability distribution.

In this case, we assume that the fraction of of buildings that are of type t are uncertain and distributed according to the beta distribution. The expert's judgment of that fraction is taken as the mean of the distribution, and the standard deviation is calculated as if the expert's judgment were based on a sample of 50 actual buildings. We then gather evidence: field observations the buildings types for a set of buildings in the community. We use Bayes' theorem to estimate what is called the "posterior distribution," the distribution of building types in the community based on the expert's judgment and updated using the observations. All the calculations can be encoded in software.

A.3 Sample Size

A.3.1 How the binomial distribution applies here

If we denote by *p* the fraction of buildings that are of type *t*, the chance of observing at least *y* specimens of a building of type t among *N* buildings that we observe is given by

$$P\left[Y_t \geq y \middle| N\right] = 1 - \sum_{m=0}^{y} {}_N C_m \cdot p^m \cdot (1-p)^{N-m} \tag{4}$$

where ${}_N C_m$ is read as "*N* choose *m*," meaning the number of ways one can choose *m* specimens of a given type among *N* samples. It is calculated as

$${}_N C_m = \frac{N!}{m! \cdot (N-m)!} \tag{5}$$

where *x*! = *x*·(*x*-1)·(*x*-2)· ... ·2·1, and *N* and *m* are nonnegative integers *and N* ≥ *m*. For reference, the summands in Equation (4) represent the binomial distribution evaluated at *m* successes in *N* trials. Note that Equation (4) assumes that buildings of each type *t* are randomly distributed through the zone, meaning for example that any arbitrarily selected building has likelihood *p* of being type t, regardless of what types of buildings are nearby.

A.3.2. At least one specimen of type t

Equations (4) and (5) are fairly abstract. To be more concrete, suppose we want to observe at least one specimen of type t among *N* samples. if the fraction of buildings (by count) that are of type t is *p*, and we observe *N* sample buildings in a zone, then the probability of observing at least one building of type t in *N* samples ($Y_t \geq 1$) is given by

$$P\left[Y_t \geq 1 \middle| N\right] = 1 - (1-p)^N \tag{6}$$

where Y_t denotes the number of buildings observed that are of type t, and $P[Y_t \geq 1|N]$ is the probability just mentioned. We can plot $P[Y_y \geq 1|N]$ as a function of *N* and *p* to help choose a reasonable sample size. The plot is shown in Figure 1. It shows that, with a sample size of *N* = 30 (3 clusters of 10 each), we are more likely than not to observe any building type that appears with a frequency of about 0.03 or more. We have a 90% chance of observing any building type that appears with frequency 0.08 or more, and 95% chance of observing any type that represents 1 building in 10. By contrast, with *N* = 9 (3 clusters of 3), we have almost an even chance of missing any building type that appears with frequency of less than 1 in 10.

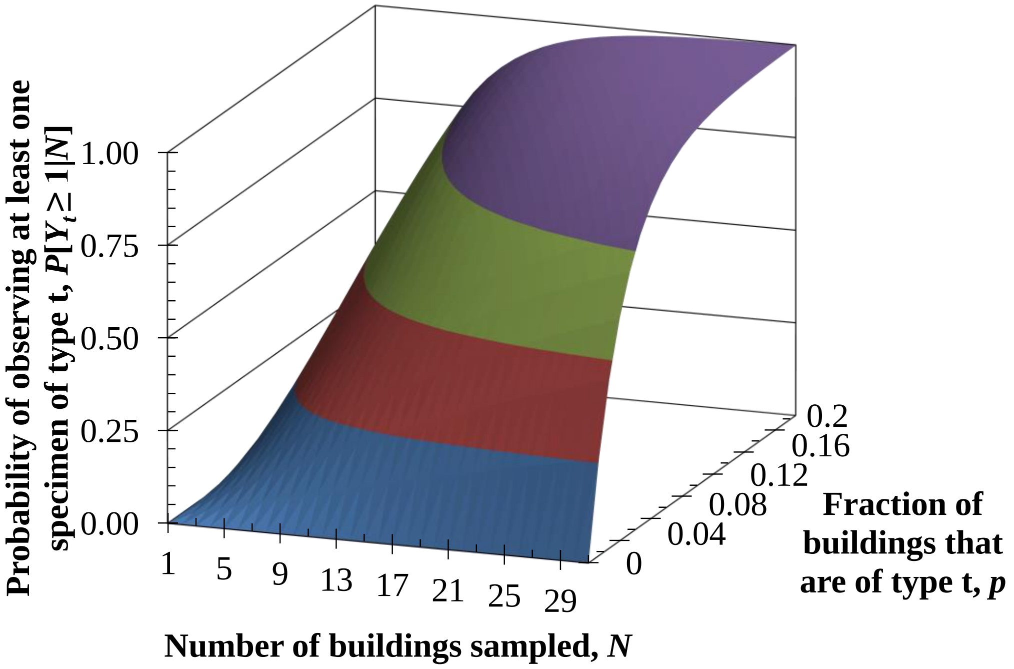


Figure 1. Likelihood of observing a building of type t, as a function of sample size and frequency among the population

So minimum sample size depends on how confident we wish to be of observing a type that appears with a specified frequency. Table 2 presents these probabilities in tabular form. For example, if we want to have better than even odds of observing a type type with frequency *p* = 0.05 (i.e., 1 building in 20), then we need to observe at least 15 of them. (See the row labeled *N* = 15 and the column labeled *p* = 0.05. It has a value of 54%, meaning 54% probability of observing at least one such building.)

A.3.2 Want to observe variability within type t

Suppose we want more than a single sample of a rare type, say $Y_t \geq 3$ samples, to be sure we know something about their variability. The probability of observing at least 3 specimens of type t among *N* samples is given by:

$$P\left[Y_t \geq 3 \middle| N\right] = 1 - \sum_{m=0}^{2} {}_N C_m \cdot p^m \cdot (1-p)^{N-m} \qquad (7)$$

Table 3 provides the probability of observing at least 3 specimens of type t among N buildings, given that the type appears with frequency *p*. With *N* = 30, we have slightly better than even odds of observing at least 3 samples of a type that represents 1 building in 10.

If we want to observe more specimens of a rare type, the probability of observing at least *y* specimens of type *t* among *N* samples is given by:

$$P\left[Y_t \geq y \middle| N\right] = 1 - \sum_{m=0}^{y-1} {}_N C_m \cdot p^m \cdot (1-p)^{N-m} \qquad (8)$$