

Spatial aspects of building and population exposure data and their implications for global earthquake exposure modeling

F. Dell'Acqua · P. Gamba · K. Jaiswal

Received: 20 January 2012 / Accepted: 28 May 2012 / Published online: 14 June 2012
© Springer Science+Business Media B.V. 2012

Abstract This paper discusses spatial aspects of the global exposure dataset and mapping needs for earthquake risk assessment. We discuss this in the context of development of a Global Exposure Database for the Global Earthquake Model (GED4GEM), which requires compilation of a multi-scale inventory of assets at risk, for example, buildings, populations, and economic exposure. After defining the relevant spatial and geographic scales of interest, different procedures are proposed to disaggregate coarse-resolution data, to map them, and if necessary to infer missing data by using proxies. We discuss the advantages and limitations of these methodologies and detail the potentials of utilizing remote-sensing data. The latter is used especially to homogenize an existing coarser dataset and, where possible, replace it with detailed information extracted from remote sensing using the built-up indicators for different environments. Present research shows that the spatial aspects of earthquake risk computation are tightly connected with the availability of datasets of the resolution necessary for producing sufficiently detailed exposure. The global exposure database designed by the GED4GEM project is able to manage datasets and queries of multiple spatial scales.

Keywords Exposure · Global database · Remote sensing

1 Introduction

Building stock quantification, their spatial location, and structural characterization are resource intensive and strenuous problems. In the real world, building characteristics such as the size, shape, configuration, architecture, material strength, structural strength, and stiffness can help describe their uniqueness from others. To systematically characterize the structural system, it is essential to compile the data related to attributes/features describing

F. Dell'Acqua · P. Gamba (✉)
Department of Electronics, University of Pavia, Pavia, Italy
e-mail: paolo.gamba@unipv.it

K. Jaiswal
U.S. Geological Survey (contracted with Synergetics Inc.), Golden, CO, USA

one or many of these characteristics, which ultimately helps to describe their susceptibility to damage and losses from different perils.

Detailed characterization identifying specific attributes of each and every building in a given region or within a country is seldom practicable given the limited resources and other constraints. Moreover, there are additional complexities involved in compiling these datasets either remotely or via field survey exercises. These are often due to limited accessibility to structural drawings, lack of familiarity of construction practice (in a specific region, or during certain timeframe), limited access/entry to interior of the building due to privacy issues, and judgments involved in the attribute interpretation process. In addition, the potential applications of these detailed datasets may be limited given the state-of-the-art of regional risk/loss estimation capabilities. Due to these practical challenges, the engineering and loss modeling community often relies on regional loss estimation strategies making use of aggregated statistical data on building stock and ‘archetype-based’ structural vulnerability models to offer realistic solutions to the global loss estimation problem.

Nonetheless, it is crucial to compile the inventory of building stock with structural characteristics and population distribution in order to perform meaningful risk and loss calculations for multiple hazard types. This article offers insights into some of the existing problems related to population and building exposure data compilation efforts and proposes practical solutions and strategies to achieve these objectives.

Existing global datasets about built-up areas (Gamba and Herold 2009) and statistical information about population and buildings in countries or regional areas (e.g., national census surveys or surveys conducted by UN agencies) are primary datasets to characterize the building stock at the global level. However, significant efforts are needed in processing these “raw” datasets in terms of their structural and occupancy characteristics. The only dataset that provides systematic compilation of global dwelling stock distribution and their structural characteristics is the PAGER (prompt assessment of global earthquakes for response) database (Jaiswal and Wald 2008), the details of which are discussed subsequently. However, the PAGER database provides inventory characteristics at a much coarser resolution than what is currently needed within the global earthquake model (GEM) application. The level of detail and the resolution necessary to perform structural vulnerability and risk computation at a global scale are indeed significant.

Multiple challenges exist to move beyond the PAGER efforts. For example, the detailed building inventory and exposure compilations are expensive and labor-intensive endeavors. Second, the public data sources that capture the essence of present needs are either incomplete or remain dispersed among multiple agencies and are often compiled in different data formats. Ancillary datasets must be identified and used in order to produce the detailed exposure information that is necessary for engineering risk applications.

An important source of building inventory data is the various national census datasets on housing. Statistical offices of countries conduct census surveys on housing/dwelling characteristics. Census surveys are rigorous, extensive (in their coverage), and sufficiently detailed (in terms of type of data that is included), unlike any other dataset at a national level. However, such an exercise is generally carried out as a part of a country’s decennial survey effort, a timeframe that may be reasonable for such a labor-intensive country-level effort, but is unable to capture the variability of building construction practices of fast-growing megacities of the world. Moreover, since structural vulnerability and risk assessment are not the central objectives for compiling such data, the level of detail is still insufficient. This is evident from the fact that often the only structural indicators compiled during the census surveys are age, construction material, and exterior characteristics.

Statistical agencies lack adequate resources/expertise to produce detailed housing data (e.g., data on housing condition, structural attributes, engineering details/durability, and secure tenure).

Detailed structural inventory datasets are vastly important. They not only facilitate rapid building damage and loss estimation but also help in understanding the impact of a housing crisis on people's lives following natural disasters. Despite their usefulness, such details have rarely been included as a part of the established data collection processes. For instance, census information is generally collected at household level instead of building level. Tenure type or wall/roof material type is included at the same level, and additional assumptions are necessary before this information can be useful for engineering risk assessments. Similarly, existing urban planning practice does not rely much on census or housing survey datasets in order to identify or locate the areas of high-risk concentration.

As a matter of fact, data measuring the internal spatial structure of the city are seldom collected in many parts of the world. Instead, the urban areas are delineated based on socio-economic information about the population, geophysical characteristics of the territories, water bodies, and artificial distinctions, such as land-use classes (with very different legends). Most of this information, including the more precise land-use data, remains insufficient for the assessment of structural vulnerability.

The lack of structural information required to evaluate building-level exposure is typically addressed through methods relying on aggregate exposure characteristics obtained from the basic parameters of the census data at coarser resolution. These basic parameters are refined using the building-level data from cadastral maps or data obtained from remote sensing imagery. The latter are used to extract building density and size, or a profile of building heights in a region. For instance, the Global Urban Observatory of the United Nations has introduced and developed a methodology that aims to estimate the durability of housing. The method incorporates information about locations that are considered hazardous (e.g., high-industrial pollution zones and high-risk concentration zones), durability of building materials, and compliance with local building codes.

For exposure, the above-mentioned methods have been designed with very different mapping schemes in order to extract the exposure information for each building or group of buildings. The process relies on inference mostly from terrestrial and remotely sensed data, and it is designed with a higher level of flexibility with respect to geographic scales and locations. In general, it may be feasible to provide very coarse statistical distributions at the regional level, but obtaining building age, size, shape, materials, and structural characteristics requires a drastically different data structure that can work at regional levels. That's why, short of an exhaustive worldwide inventory of structures, a statistical approach for characterizing structural classes based on auxiliary datasets will be required to develop a consistent and reasonably accurate global exposure database (GED).

Moreover, there is also a need to integrate existing inventory and vulnerability databases within the GED. These databases are from different parts of the world and often are of different resolutions, for example, town, district/county, and/or lower resolutions. Absent clear guidelines for exposure compilations at detailed and aggregated levels and absent effective coordination among the stakeholders, it is difficult to leverage the ongoing regional efforts in order to produce a high-quality, global, multi-scale, and regionally/locally driven exposure database.

The background tasks necessary before fully conceptualizing the data structure of GED for GEM are as follows:

- (a) Analysis of the state-of-the-art methodologies for mapping the data into relevant structural typologies including estimation of building occupancy loads, estimation of economic exposure, and possibly integration/validation of these with remotely sensed data/inputs.
- (b) Review of existing global datasets (that are available or eventually may become available to GEM), and development of processes to incorporate such datasets.
- (c) Exploration of the potential for using on-going research/commercial developments in the field so as to produce a more reliable and useful exposure database.

For the first task, the most relevant example is the Federal Emergency Management Agency's multi-hazard loss estimation software, widely known as HAZUS (Schneider and Schauer 2006). HAZUS uses a "mapping schemes" approach to distribute square footage by occupancy into structural classes that are directly linked to damage functions. These mapping schemes capture the distribution of specific occupancy categories (e.g., multi-family residential occupancy) into model building types (defined as a part of HAZUS taxonomy) that include distinctions such as structural details, height (defined in terms of number of stories), era (to define design code levels), and quality of construction. The default mapping schemes delivered within the HAZUS platform are crude and are designed to cover multiple states due to the lack of additional data to infer separate mappings. As the program matured, several region-specific studies/implementation projects improved the default mapping schemes for specific areas, largely through a process of analyzing tax assessor databases or interviewing engineers in public works. Currently, the San Francisco bay area, the counties in Southern California, the New York metropolitan area, and the State of South Carolina have custom mapping schemes. These detailed mapping schemes are available for subsequent risk studies, enabling the user community to harness work done by others to characterize the structural distribution of building stock.

Although a few projects have adapted the HAZUS platform for international use, the mapping schemes were designed and tailored for US-specific application because the underlying database architecture is dependent on the US Census data and HAZUS structural taxonomy scheme.

As a natural extension of the ShakeMap program (Wald et al. 2005), the U.S. Geological Survey (USGS) has recently developed a rapid post-earthquake impact estimation system called PAGER (Prompt Assessment of Global Earthquakes for Response). Unlike HAZUS, the PAGER system works internationally. The USGS PAGER program adopts an approach similar to the HAZUS mapping schemes, but the team has developed a wide variety of schemes that depend on the nature and quality of the available datasets needed for structural categorization and creating global building inventories (Jaiswal and Wald 2008). However, due to the lack of detailed land-use information at the worldwide level, the team relied on CIESIN (Center for International Earth Science Information Network)'s Global Rural Urban Mapping Project (GRUMP) dataset (Salvatore et al. 2005). It helped to distinguish between the urban and rural areas in any given country, so that the structural variations of residential and non-residential buildings in these settings can be systematically accounted (Jaiswal and Wald 2010). The authors have analyzed several datasets that include the country-specific aggregate census data, the demographic data from the United Nations Statistics Division, and the EERI World Housing Encyclopaedia (Brzev et al. 2004) in order to characterize the building stock. For 247 regions and countries, the USGS developed statistical profiles of residential and non-residential structure types in both urban and rural environments. The mapping schemes resulted in a structural inventory dataset that varies in quality from country to country, and is limited by the spatial

resolution of the information available. A significant challenge is that the “mapping schemes” that provide the basis for distributing building proxy data into structural classes are dependent upon types of input datasets. However, these schemes can vary greatly throughout the world as newer data become available, thus newer mapping schemes must be developed.

The USGS PAGER system addresses building exposure at a global scale, providing a critical starting point and significant progress toward the envisioned GED. The PAGER/HAZUS endeavor incorporates best practices critical to the creation of a global dataset, including statistical characterization of structural type by country, leveraging of existing global datasets, and proven methods of extrapolating structural parameters from UN housing and demographic data for the purpose of estimating losses. It is anticipated that PAGER’s current structural taxonomy (PAGER-STR is discussed in Jaiswal and Wald 2008) will be a starting point for the ontology and taxonomy to be established in the next phase of the GEM project. However, GEM will require additional structural types and classes and corresponding vulnerability functions that extend beyond this work.

The state-of-the-art on global human settlement datasets and associated procedures is quickly evolving, as documented in the analysis of Gamba and Herold (2009). These datasets are largely based on moderate resolution optical data, but they are also likely to be extended to high resolution in the near future (Pesaresi et al. 2008), as well as to radar data (Dell’Acqua 2009). In certain research initiatives, the joint use of radar and optical datasets seems to be a positive and encouraging development (Polli et al. 2009). Moreover, the spatial analysis can no longer be considered at a single scale because information about human settlements in recent times is becoming available at much higher spatial resolution than its predecessors. It is likely that future datasets may consist of varying resolutions depending on the importance/land cover/land use of the area. In urban areas with dense population, more details may be necessary for inference about the spatial characteristics of building stock exposure. However, rural areas are likely to be imaged with coarser details and will probably demand different procedures to integrate data with separate mapping schemes. Schneiderbauer (2007) demonstrates population mapping aimed at computing the social vulnerability component of the risk equation. Datasets like the Global Risk Data Platform (Giuliani and Peduzzi 2011) developed by UNEP (United Nations Environment Programme) are thus likely to remain too coarse for sophisticated risk analyses, although they are a very important source of information.

Examples of proprietary databases include those developed by commercial and reinsurance companies. Such datasets are designed to take into account as much exposure-related information as possible of varying resolution. For instance, Swiss Re’s Reinsurance and Risk Division is developing a tool called CatNet (Schmidt 2001), which is an on-line electronic atlas designed to capture a range of information at different resolution on natural hazards at the global level. Similarly, risk management solutions (RMS)’s RiskLinkTM software provides earthquake exposure information at aggregated levels such as CRESTA-level as well as detailed levels based on address-specific exposure classification for risk analysis. But despite their coverage and quality, it is highly unlikely that these datasets or the associated procedures will be available for GED creation due to their proprietary nature.

2 A global exposure database for GEM

Within the challenging arena of the GEM, the GED4GEM project (GED4GEM 2011) aims at developing a global exposure database, or GED, an effort central to assessing earthquake

vulnerability and risk assessment worldwide. The aim of the GEM project is to make a collaborative effort to develop and deploy tools and resources for earthquake risk assessment worldwide. The earthquake risk can be computed in terms of probabilistic estimation of likely economic and human losses from future earthquakes anywhere in the world. In doing so, GEM users may also be interested in computing estimates of the building damage and collapses in a given area affected by a scenario seismic event, and/or estimating the associated casualties or the earthquake's impact on socio-economic recovery.

The biggest challenge for GEM computational facility is therefore to serve these objectives at a global scale that requires development of underlying datasets, models, and computational tools and architecture. The key steps consist of performing earthquake hazard analyses (estimating probability of certain levels of shaking), exposure analyses (estimating building and population characteristics at a specific location or in a particular area), and vulnerability and loss analyses (estimating damageability and losses associated with underlying hazard and exposure condition) for a chosen location or specific region. The GED will be the main source for exposure data and thus a central element of the whole GEM architecture.

Due to the general nature of the exposure information in the GED, the outcome of the GED4GEM project may also be useful for performing multiple risk and loss computations, for example, estimating economic losses, and casualties due to floods (Merz et al. 2007), landslides (Glade et al. 2005), hurricanes (Waters et al. 2010), tsunamis (Woods 2009), and volcanic eruptions (Lirer et al. 2010).

According to the general aims of GEM and GED4GEM, the structure of the project aims at achieving the following three tasks:

1. Identification, collection, and homogenization of global datasets that may be accessible and are useful for the GED development.
2. Definition and implementation of sets of procedures to build a GED based on the output of task 1.
3. Definition of “best practices” aimed at populating missing attributes and/or layers of information in specific geographic areas.

These tasks will be conducted in two distinct phases. The first phase consists of identification and collection of existing databases, mainly to carry out in situ as well as web-based surveys. This phase began in November 2010 and was completed within the first 9 months of the project (Huyck et al. 2011). In essence, this phase has allowed identifying and including most worldwide publicly available building inventory datasets. In addition, other datasets were identified; however, they required newer inventory and structural occupancy mapping schemes.

The second phase mostly constitutes compilation and harmonization of data using “best practices”, and development of the procedures for its management to allow updating and further improvements. The insertion of existing databases into the new GED should be a task easy to perform via semi-automatic approaches. It consists of implementing mapping schemes developed by the Inventory Capture Tools consortium (see the section on BREC4GEM). The second phase covers the rest of the time span of the project, which is expected to be completed by the end of 2013.

The main features of the GED4GEM output will therefore be as follows:

- A database structure that is compliant with the Open Geospatial Consortium recommendations and is also consistent with the guidelines and requirements of other projects running under the GEM framework.

- Semi-automatic algorithms that enable data ingestion, future updating, etc. The procedures allow the fusion of new and alternative sources of data in order to obtain geographically coarse yet statistically consistent earthquake exposure for risk assessment.

The “harmonization” of the different data sources will require devising a new database structure that allows reconciliation of the available datasets. Links among spatial databases with different resolutions and structures will be implemented as much as possible, and standards-based tools will also be developed to support access, querying capabilities, and visualization functions. Data fusion algorithms will be considered to infer the missing data and other information necessary for GEM’s ongoing needs.

2.1 Which spatial scales are relevant to the GED development?

The above-mentioned objectives imply that spatial scales are relevant to creation of the GED in two different ways:

- (a) First, determinations of the level of spatial resolution at which consistent statistical information about the assets (i.e., population and buildings) exists.
- (b) Second, determination of the spatial scales that are relevant in order to (dis)aggregate existing data and then eventually produce user outputs at different spatial scales if necessary via a web interface.

For input statistical data, consistency depends on the source. Since such data are usually obtained from surveys (e.g., national census surveys or regional surveys administered by local governments), they are consistent at the national, sub-national, town, or census tract level. For output statistical data, the GED is expected to provide information about population and buildings with a spatial resolution encompassing the input information after homogenization among different countries, regions, or sub-national areas. Therefore, different input scales are always provided to the final users of the gridded (raster) system. Specifically, the GED uses a 30 arc-second grid (roughly corresponding to a 1-square-km spatial resolution) to store and support queries related to the processed data.

However, there are certain data sources (e.g., street-level surveys or tax assessor’s data), where building-level assessments are provided. This demands data ingestion at a much finer spatial resolution, and it is not compatible with statistical datasets that employ raster-based resolution. This configuration requires an alternate way to store and distribute the data, by means of a vector-level geographic information system (GIS) representation.

Accordingly, the spatial scales of the inputs and outputs of the GED resulted in four different spatial scales (labeled as “levels”) in the final database:

- “Level 0”: a gridded dataset, whose values are spatially disaggregated at 30 arc-second grids, and whose statistical consistency is ensured at the coarser spatial scale, corresponding to countries or regional levels, for example, data compiled through national census surveys.
- “Level 1”: a gridded dataset, with the same grid resolution as the previous level, but with inputs coming from sub-national or district-level surveys to the disaggregation schemes. The statistical consistency is ensured at a finer spatial scale, for example, at the county level in the USA, or at the *arrondissements* level in France.
- “Level 2”: a gridded dataset with statistical consistency based on datasets compiled at each grid level, obtained by aggregating sub-grid/finer data, such as those coming from building-by-building surveys.

- “Level 3”: a vector GIS dataset that includes single building/dwelling-based compilations, located with address-level precision and with detailed structure-specific attributes.

3 Spatial disaggregation of population counts according to building statistics—the PAGER paradigm

As mentioned above, the PAGER inventory database is the most detailed publicly available dataset. It was compiled using a variety of sources including (i) country-specific housing census surveys, (ii) UN and UN-Habitat inventory databases, (iii) the World Housing Encyclopedia (WHE) database developed by the Earthquake Engineering Research Institute (EERI), and (iv) data obtained through WHE-PAGER project by conducting expert judgment surveys. The PAGER inventory database provides the distribution of housing/dwelling units rather than the distribution of buildings by urban-residential, urban-non-residential, rural-residential, and rural-non-residential occupancy types. In most census databases, the housing/dwelling units represent independent abodes for a single household/family. In general, the dwelling-type distributions can be considered as a proxy for population distribution by structure types; however, the same is not true in the case of building distributions by structure type. This is mainly because several single-family or multi-family dwellings may be housed within a single building and the number of dwelling units within a building may vary from building to building. The PAGER inventory approach relies on estimation of indoor population within different construction types at different times of the day (e.g., day, night, or transit hours). Details about inventory database development are discussed in Jaiswal and Wald (2008, 2010).

PAGER’s structural inventory methodology can be directly used to develop the Level 0 exposure layer for population within the GED4GEM database. Lacking information about actual building counts, as further discussed in next section, the PAGER database needs to be complemented by another approach to extract building information on the same 30'' by 30'' grid. Figure 1 shows the distribution of population counts in wooden houses at 10 a.m. in Japan shown in terms of a colored map where the people counts are graphically represented by a reddish color ranging from black—minimum value—to bright red—maximum value.

The next step in the same procedure, which includes spatial enhancement of the population classes within the GED grid cells, is the evaluation of different statistical distributions for building stocks according to sub-national areas, as described in the “Level 1” layer of the GED. We illustrate this using Bolivia’s sub-national building stock data that were obtained from the 2008 Demographic and Health Survey (DHS) study (Fig. 2). As long as there are enough samples in a given sub-national area to provide a statistically robust distribution, this approach allows a simplified, although spatially discrete, way to make the population disaggregation better fit the local building distribution, without altering the national statistics, thus ensuring a complete back-compatibility to “Level 0” data.

4 Spatial disaggregation of building counts according to population—the GED4GEM paradigm

Despite its wider coverage, the PAGER paradigm is not immediately useful for building asset characterization in terms of number of buildings per class. This is due to the fact that the

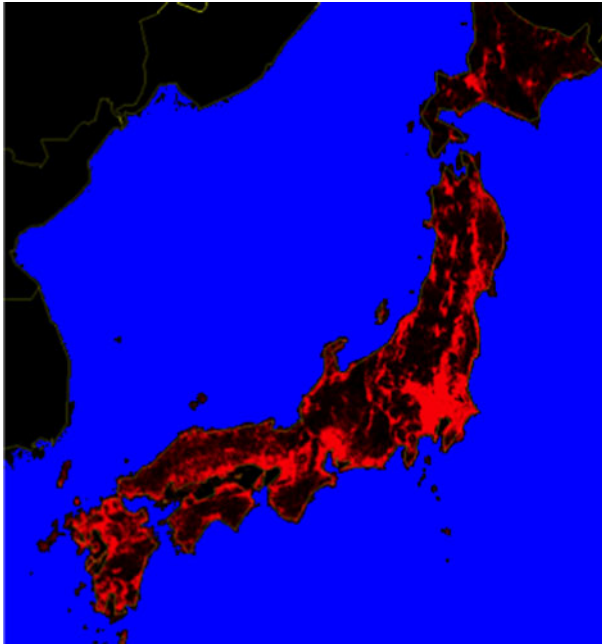


Fig. 1 Graphical representation of the Japanese population within wooden houses at 10 a.m.: The spatial distribution follows the population data (2010 LandscanTM data were used for this simulation), but the numbers in each cell (represented in *colors*) are decreased according to a percentage evaluated using the overall percentage of wooden houses compared to other construction types in Japan. Cells in other countries are left intentionally with null values (*black*)

PAGER inventory methodology was developed to rapidly estimate human occupancies in different construction types. The objective was to estimate casualties within an order of magnitude accuracy immediately following a large earthquake anywhere in the world. Due to the lack of open, publicly accessible datasets with sub-national resolution, the authors developed procedures to approximately infer building stock characteristics, indoor population, and other spatial characteristics. It was assumed that the country-level statistical distribution of different building types is representative of the sub-national structural type distribution. Specifically, by assuming only urban or rural building stock distributions, the PAGER approach avoided the need to count the actual number of buildings; instead, it deduced a distribution of population according to different building types, a step necessary for human casualty estimation. This was an efficient way to expedite the casualty loss estimation within the PAGER system given the lack of data. However, due to the unavailability of actual building counts or information about floor areas of different building types, it is difficult to compute building damage and economic losses. In order to overcome this difficulty, at least partially, the GED4GEM project aims to develop a complementary approach that can enable estimation of the missing attributes. Specifically, GED4GEM includes additional structural and demographic information in the loop as discussed in subsequent section.

4.1 Past approaches to the problem

Jones et al. (1976, 1987) studied the building counts, building sizes, their uses, and total population data from a number of cities around the world, and developed geographic

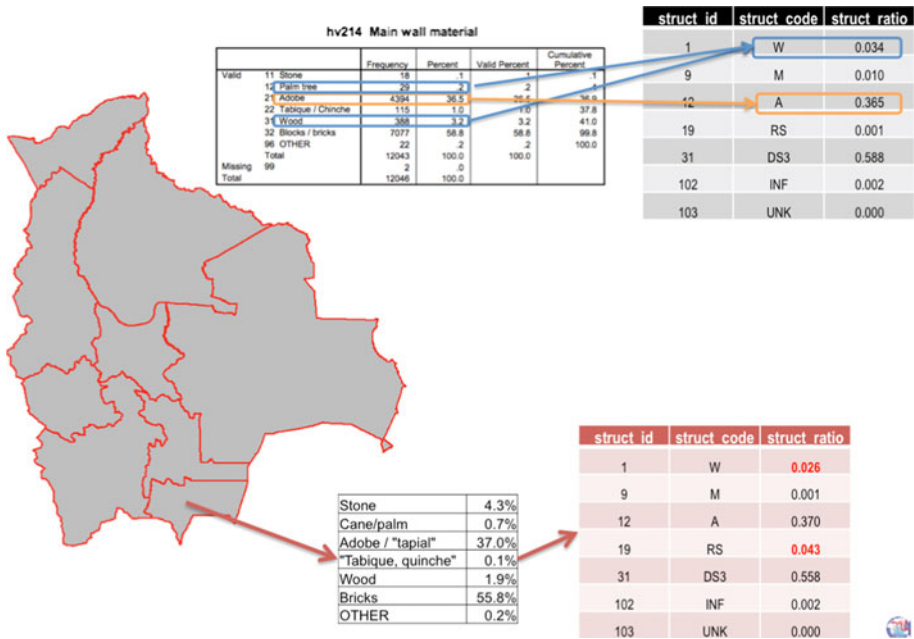


Fig. 2 Comparison of statistical distributions of building feature information extracted from DHS data for Bolivia, and the corresponding building distribution according to PAGER taxonomy (“W” wood, “A” adobe block, “M” mud wall, “RS” rubble stone masonry, “DS3” dressed stone or block masonry with cement mortar, “INF” informal/makeshift construction)

location-dependent relationships between total population and several building-specific parameters. For example, Jones et al. (1976) developed indirect relationships between total number of buildings and total population by combining data from various cities in various years for multiple countries. Similarly, Jones and Lewis (1990) suggested a simple non-linear relationship to predict the number of buildings of different size classes according to the following law:

$$E(\ln PB_i | \ln M_i) = \alpha + \beta \ln U_i. \tag{1}$$

where PB_i denotes the percentage of total buildings greater than size class i , and U_i indicates the upper bound value of building size class i . The terms α and β are regression parameters that must be derived using sample datasets. Building sizes are expressed in square footage and were categorized into a number of intervals/classes.

Jones and Nicolaidis (1988) showed that such relationships are stable and hence it is possible to use such relations at places where building count data are lacking. The authors noted that many variables could influence these relationships (e.g., per capita income, culture, climate) and envisioned greater reliability if such relations were applied to large geographic regions than to smaller sub-areas such as specific municipalities. In earlier efforts, such relationships were explored only for specific areas of studies. Leach (2001) recently performed a multiple regression analysis on U.S. census data. Using the residential block data available for three cities, the author developed empirical relationships to estimate the mean values of building counts, building sizes, and average building heights at block levels. Among the three regression models, the model on building count estimation

per square miles appears promising, with a limited predicting power. Thus, the regression models by themselves cannot offer a complete solution to GED4GEM project. While developing such relationships, care must be taken in selecting a study region, in identifying potential influencing parameters, and finally in applying such relationships only to representative geographic settings.

4.2 The GED4GEM approach

As discussed earlier, the GED4GEM database aims to provide grid-based earthquake exposure derived from input datasets that are available at multiple resolutions. Since the primary source of global exposure information is gridded population datasets, we needed to develop empirical relationships in order to produce region-specific estimates of total building counts and their sizes using regional population data and auxiliary information.

The following steps are essential in devising a strategy to derive building counts from different types of input data:

- *Estimating building counts using population data:* As discussed earlier, the goal is to derive region-specific or individual setting-specific (such as rural, semi-urban, urban) empirical relationships that help estimate total building counts (irrespective of their use types) directly from underlying population datasets.
- *Estimating the number of housing units using average occupancy/household data:* In addition to population data, the national statistical organizations also compile data on housing characteristics such as material of construction, average size of household, number of households, tenure, and number of occupants. Huyck et al. (2011) provide the list of countries for which Demographic and Health Survey (DHS) data are available within UN-HABITAT records. By estimating the average number of occupants per housing unit in a specific study area, it is possible to approximately estimate the total number of housing units directly from population data.
- *Building counts from sampling and scaling:* This step requires more extensive ground surveys than those cited in the previous steps. Rigorous ground-truth surveys can be conducted in specific study regions in order to compile sample datasets and to develop statistical relationships for further application beyond the study areas. Alternatively, the same information can also be extracted using remote-sensing data (see the section about BREC-4-GEM for a more detailed description of the methodology). Once enough samples are compiled characterizing specific built environments, it is possible to infer an approximate number of people per building and then use such rates to compute the total number of buildings for any new area that is analogous to the study area. Similarly in the case of the remote-sensing application mentioned earlier, it is also possible to establish a relationship between a sample study area and the total building count and then use it to estimate the total number of buildings of a new region with a similar built environment. For example, if A_{Z_i} represent the total geographic area (in sq km) covered by the i th study zone Z_i and the term b_{Z_i} represents the total buildings that are accounted for within that zone, then the total number of buildings in any new region R with geographic area A_R can be estimated as $b_R \approx E(b_{Z_i} \cdot (A_R/A_{Z_i}))$, where the expectation is approximated by the mean over all the study regions Z_i . It is possible to refine this approach further by considering different occupancy type zones (e.g., high-rise commercial occupancy zone in Central Business District—CBD, low-rise residential occupancy type in suburban areas, etc.). Land-use maps prepared by national/regional mapping agencies or maps extracted using remote-sensing data can be

used as a starting zonation model to apply such schemes. A similar procedure was adopted under the NERIES (Network of Research Infrastructures for European Seismology) project to compile building inventory data for many European countries and Turkey (see Erdik et al. 2010 for additional details).

- *Building counts from cadasters or national inventories*: Some of the countries have already started conducting building-level census surveys instead of population or household-centric census surveys. Building-level census surveys directly provide total building count, and it is the best possible raw data that can be directly assimilated within the GED4GEM database. Similarly, certain countries with national cadasters or pre-compiled building inventories (compiled for tax purposes) can be directly incorporated within the GED. However, building-level details are seldom made available for public consumption because of privacy concerns. Aggregated data from national agencies/government may be available, but cost may be another impeding factor for achieving significant global coverage.
- Among the three approaches discussed above, the one that may be more feasible to apply at the country and sub-country level (i.e., at levels 0 and 1) is that of extracting the building counts at the grid level (i, j) using the population dataset. Instead, a more complex relationship $b(i, j) = f(p(i, j))$, such as those suggested by Jones et al. (1976), may be estimated and directly applied only if enough building count samples are available for a specific geographic area (e.g., a country) to derive region-specific regression parameters. It is difficult to find a sufficient number of samples for every country in the world; thus, this approach will be limited to the finer (“Level 2”) spatial scale.

In order to formalize the proposed methodology, let us define a few basic quantities, whose values may be available at different scales (national or sub-national):

1. *Average number of people per dwelling* (\bar{p}_d): This information is accessible from most census datasets; however, it is usually published as an aggregate estimation and does not distinguish among different structural typologies (such as concrete, masonry, or wood-frame buildings);
2. *Average number of dwellings per building type* ($\bar{d}_b^{(k)}$): Even though such information is not readily available, it can be assessed approximately for each structural/building type (as indicated by the index k) using detailed ground-truth surveys for a given area, or can be inferred from judgments drawn from expertise in engineering and construction practice;
3. *Average number of people per building* (\bar{p}_b): This can also be accessible from most census datasets. As an alternative, the ratio between the total (urban) population and the total building count in an (urban) area may be also considered.

The GED4GEM population-to-building occupancy approach uses dwelling-type fractions $df^{(k)}$, obtained (e.g., as in the PAGER’s inventory dataset) by means of statistical analysis of census data, national surveys, or existing technical literature. In order to retrieve total building counts and fractional building counts $b^{(k)}(i, j)$ for the (i, j) th element of the GED grid, the procedure works according to the following steps:

1. First of all, an estimate of the building count $b(i, j)$ is obtained by a combination of one or more of the above statistical quantities, according to what is available at the same or coarser-scale level. Specifically:

- a. If \bar{p}_b it is available and is applicable at this grid-cell resolution, estimate the total building count at each grid cell as

$$b(i, j) = p(i, j) / \bar{p}_b. \tag{2}$$

- b. If \bar{p}_b is not available but the aggregated building count b_Z for the area Z with total population $p_Z = \sum_Z p(i, j)$ is available, then $b(i, j)$ can be extracted using the local population count normalized to the aggregated building using

$$b(i, j) = \frac{p(i, j)}{p_Z} \cdot b_Z. \tag{3}$$

- c. If \bar{p}_d and \bar{d}_b^k are available, then the number of dwellings per type $d^{(k)}(i, j)$ can be estimated as $(p(i, j) \cdot df^{(k)}) / \bar{p}_d$ and then estimate the building count at each grid cell using

$$b(i, j) = \sum_k \frac{d^{(k)}(i, j)}{\bar{d}_b^{(k)}}. \tag{4}$$

- d. Finally, if only \bar{p}_d is available, then we may assume that $\bar{d}_b^{(k)}$ is approximately 1.0 for single-family/single-story buildings, 6.0 for mid-rise buildings, and 15.0 for high-rise residential buildings, and apply the algorithm as described in c.

2. Once the total building count $b(i, j)$ has been estimated at each grid cell, next step is the extraction of fractional building counts $b^{(k)}(i, j)$ for the (i, j) th element of the grid for each building type k . We use the principle described earlier:

- a. If $\bar{d}_b^{(k)}$ is available, then building-type fractions are estimated as $bf^{(k)}$

$$bf^{(k)} = \frac{b^{(k)}}{b} = \frac{d^{(k)}}{\bar{d}_b^{(k)}} \frac{1}{\sum_k (d^{(k)} / \bar{d}_b^{(k)})} = \frac{d^{(k)} / d}{\bar{d}_b^{(k)} \sum_k (d^{(k)} / \bar{d}_b^{(k)}) / d} = \frac{df^{(k)}}{\bar{d}_b^{(k)}} \frac{1}{\sum_k (df^{(k)} / \bar{d}_b^{(k)})}. \tag{5}$$

and then $b^{(k)}(i, j) = bf^{(k)} \cdot b(i, j)$.

- b. If \bar{d}_b^k is not available, the assumptions at step 1(d) can be applied in addition to Eq. (5) used to approximately estimate $bf^{(k)}$.

By applying this algorithm, and possibly assuming that building counts at the aggregated level are available, both “Level 0” and “Level 1” of the GED may be populated, although the quality of the final results may be different for various geographic regions, depending on the available data. Example results are shown in Fig. 3 using the 2005 GPW gridded population values (Balk and Yetman 2004; GPW 2004) for (a) Finland, where a total building estimate (1,504,473 in 2007) is available (Erdik et al. 2010), and (b) Saudi Arabia, where \bar{p}_d (6.1 person/dwelling in 2004) was obtained from census information available on-line.

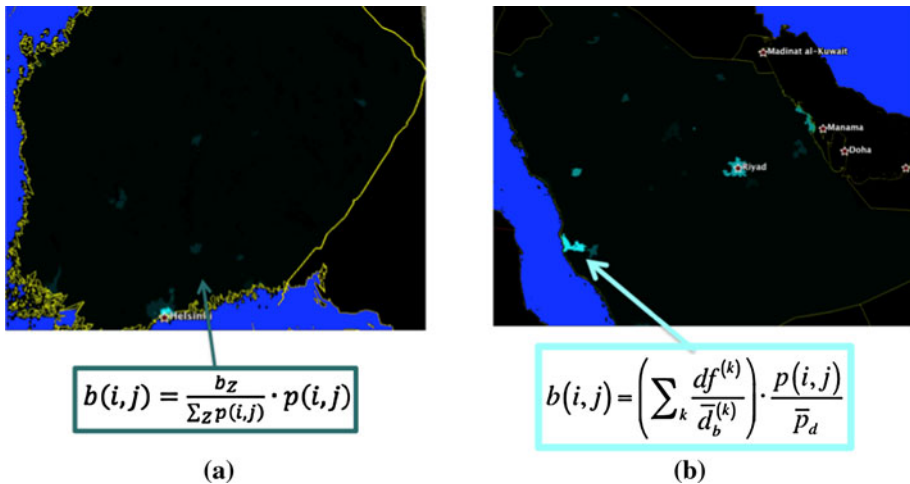


Fig. 3 Graphical representation of the number of buildings in the $30'' \times 30''$ GEM grid (level 0) for **a** Finland and **b** Saudi Arabia, as a result of the procedure described in the text to extract building counts from population. Color scale ranges from *black* (0) to *white* (maximum building density) through different color shades. Cells in other countries are left intentionally with null values (*black*)

5 BREC-4-GEM: a tool for extracting building counts from remotely sensed data

As mentioned in the previous section, building disaggregation starts from building counts in a specific area, and there is therefore a need to obtain such numbers. At the aggregated level (without a detailed building-by-building recognition of their typologies), this problem can be solved by using remotely sensed images. In fact, an increasing number of Earth Observation satellites continuously collect data about the Earth surface at increasingly finer spatial resolutions. For this purpose, the Inventory Data Capture Tools project (IDCT 2012) was started to feed information into the GED4GEM system. Specifically, the software tool BREC-4-GEM, standing for “Built-up Area RECOgnition for GEM”, has been designed to automatically extract key building inventory parameters including footprints, height, shape, and damage occurrence from very high-resolution optical imagery. BREC-4-GEM builds on BREC, an existing software suite developed over several years that can extract various elements of the urban environments (buildings, roads, land-use/land-cover classes) from either optical or radar, multi-spectral and hyperspectral satellite and airborne datasets (Gamba et al. 2009).

The need for a new tool stems from the fact that a considerable body of technical literature exists on the topic of building count feature estimation from remotely sensed data, together with several important examples of operational implementations; however, a consistent and standardized methodology or toolkit for building stock data collection for inventory purposes was still missing.

Each BREC tool to be integrated into the new software was transformed into a QGIS plug-in (Athan et al. 2012). The conceptual scheme is shown in Fig. 4.

For the purposes of this paper, the useful BREC-4-GEM plugin is the one devoted to automatic extraction of building footprints, which also allows the collection of building counts in a given area. This tool is capable of extracting much more than just footprints, for example, the building exterior/facade, the 3D shape and size of a building and height,

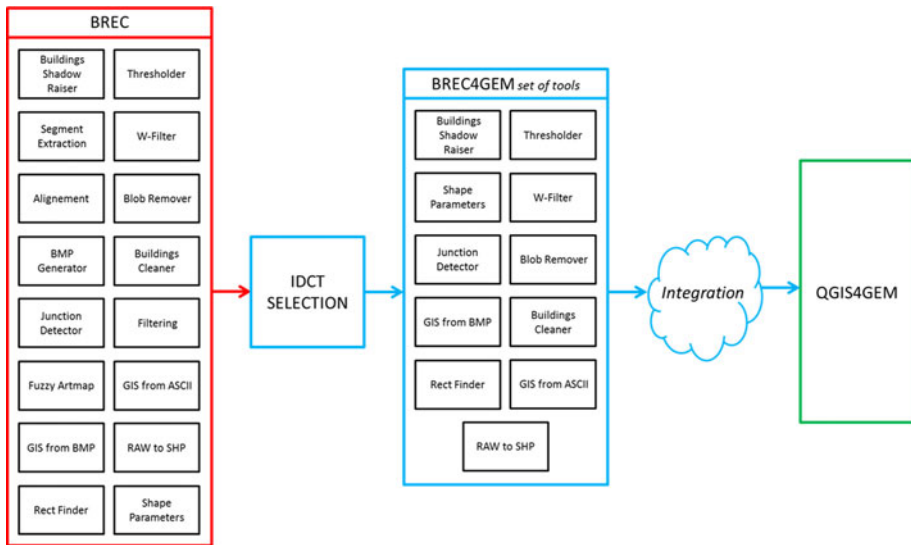


Fig. 4 A graphical representation of the process of selecting and integrating BREC tools into QGIS

which are instrumental for inferring their typologies. The study area used for illustration purposes is the first test case considered for the IDCT project, that is, a Quickbird image of the city of Pylos, Greece (Dell’Acqua et al. 2012).

For the extraction of building counts and footprints, two different procedures are available. The first uses an unsupervised radiometric classification to process the data. The second uses a supervised classification process.

Use of unsupervised/supervised classifications produces classifications like those shown in Fig. 5, where the classifications are shown side-by-side with the corresponding crop of the panchromatic images. The locations and shapes of the buildings have been essentially captured, although some artifacts appear (see the jagged appearance of the “light green” class caused by roof sections facing the sun).

In order to extract building counts and footprints, a binary (building/non-building) image is obtained by reassigning the output classes of the abovementioned classification maps to obtain black and white images, where possible “building elements” (or “blobs”) are further analyzed, according to the following procedure:

1. First, BREC-4-GEM removes small “blobs” corresponding to false-positive classifications. For this purpose, the BREC tool ‘*Blob Remover*’ is used, which deletes building pixel blobs smaller than the ‘lower threshold’ and bigger than the ‘upper threshold,’ assumed to be respectively too small and too big to be actual buildings according to usual construction practice in the test area.
2. Then, using the ‘*Building Regularizer*’, another BREC tool, the building shapes are regularized. The tool forces some regularity constraints on the building footprint, such as right angles and straight lines.

Final results of the procedure are shown in Fig. 6. It is clear that although the building details may be different and highly approximated due to imprecision of the classification step, the building count is the same for the two approaches.

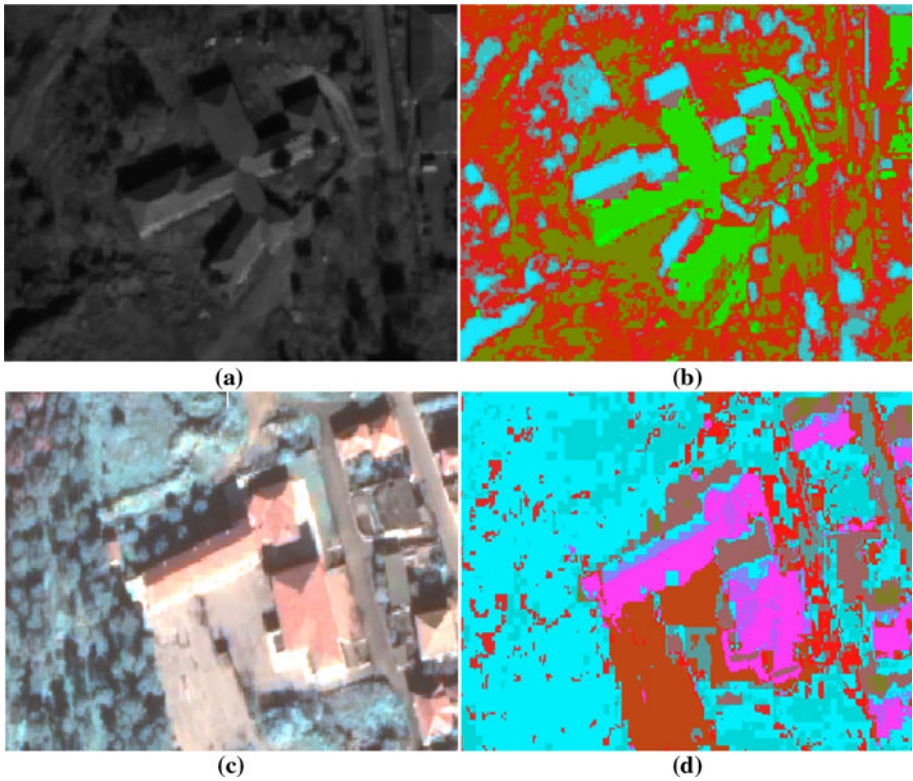


Fig. 5 **a** a small section of panchromatic image; **b** the corresponding result of 9-class unsupervised classification; **c** another section of the same panchromatic image; **d** the corresponding result of 9-class supervised classification



Fig. 6 *Red lines* define polygons with the outline of extracted buildings: **a** extraction using the unsupervised classification; **b** extraction using the supervised classification

As a general comment on the results that were obtained on a number of scenes like the one in Fig. 6, the sheer presence of buildings is detected with around 90 % accuracy. As an example, the results for the complete Pylos urban area are shown in Fig. 7, where each detected building is highlighted by a red dot.

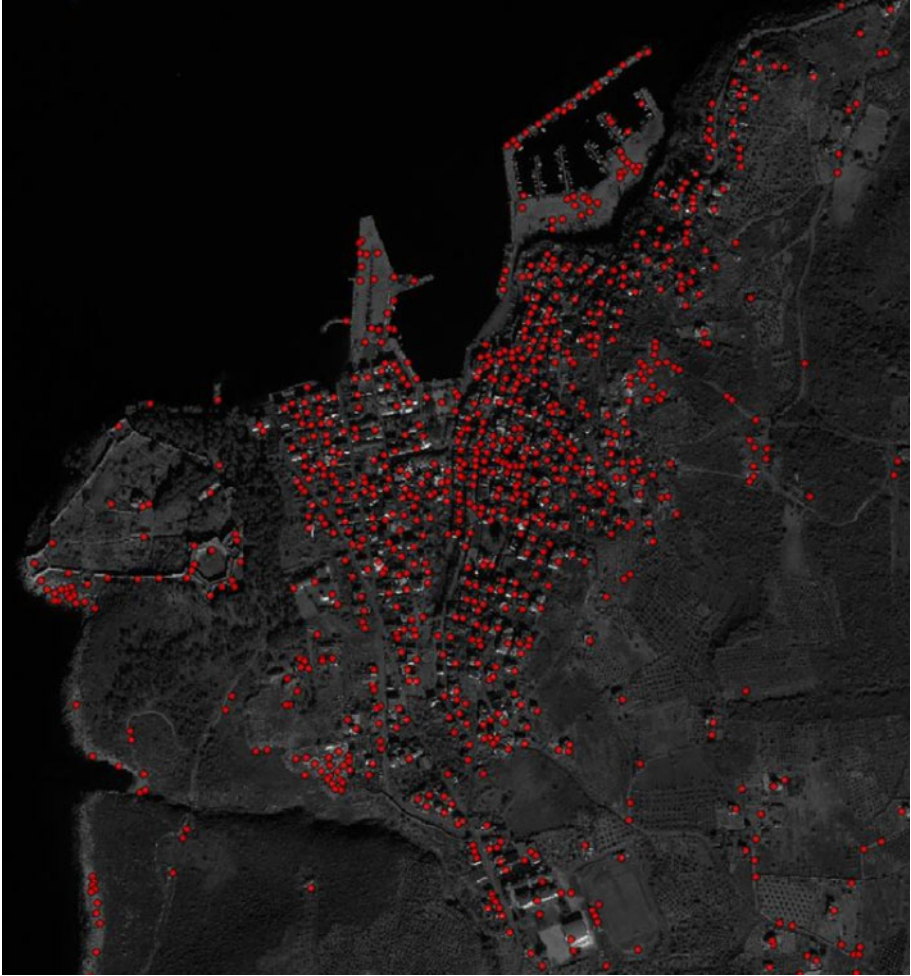


Fig. 7 Overall building extraction results for the Pylos urban area

The goal of using BREC-4-GEM in this context is the possibility of collecting building counts in multiple test areas, rather than using it as a tool to bulk process a whole country, let alone the whole world. The spatial resolution of the required remotely sensed imagery would result in a huge data volume, whose processing is not (yet) within the reach of current systems. By analyzing instead significant samples of human settlements, the BREC-4-GEM suite may provide valuable information to develop, for instance, the population-to-building models described in the previous section.

6 Conclusions

The spatial aspects of building distributions for earthquake exposure collection are a challenging part of a Global Exposure Database. The GED4GEM project attempts to

address some of these aspects via a combination of the spatial disaggregation of population data using building statistics, and the extraction of building counts using population as well as ancillary statistical data. Results and methodologies are preliminary but thrilling, as they provide new ways to look at existing information and combine them into spatially meaningful datasets. The next step is to validate the numbers obtained by using independent sources, such as ground-based surveys or cadastral data. This effort is ongoing and should produce a global and consistent evaluation. Further approaches that were tested only at the regional level, such as the one developed for Europe in the NERIES project (Erdik et al. 2010), need to be applied and tuned to other areas, the main goal for the computation of the building count in the “level 1 GED”.

Acknowledgments The research performed in this work was funded by the Global Earthquake Model (GEM) Foundation, in the framework of the “Global Exposure Database for GEM” (GED4GEM) and “Inventory Data Capture Tools” (IDCT) projects. The authors are in debt to all the other people involved in these projects for useful discussions about this topic, especially to Charles Huyck (ImageCat Inc.) and Gianni Cristian Iannelli. A special thanks to Davide Cavalca, IT manager of the GED4GEM project, for his help in managing most of the data exploited in this work. Ms. Margaret Hopper of USGS helped review the manuscript to improve its readability. Mention of trade names or commercial products does not constitute endorsement or recommendation for use by the U.S. Government.

References

- Athan T, Dassau O, Ghisla A, Homann M, Macho W, Engel A, Fischer J, Sherman F, Blazek R, Dobias M, Holl S, Carlson K, Farmer J, Morely B, Hugentobler M, Sutton T, Contreras G, Ersts P, Horning N, Luthman L, Mitchell T, Willis D, Macaulay W (2012) Quantum GIS user guide v, 1.7.0. Available online at http://download.osgeo.org/qgis/doc/manual/qgis-1.7.0_user_guide_en.pdf
- Balk D, Yetman G (2004) The global distribution of population: evaluating the gains in resolution refinement, CIESIN. Available at http://sedac.ciesin.columbia.edu/gpw/docs/gpw3_documentation_final.pdf
- Brzev S, Grenee M, Arnold C, Blondet M, Cherry S, Comartin S, D’ayala D, Farsi M, Jain S, Naem F, Pantelic J, Samant L, Sassu M (2004). The web-based world housing encyclopedia: housing construction in high seismic risk areas of the world. In: 13th world conference on earthquake engineering, Paper No. 1677
- Dell’Acqua F (2009) The role of SAR sensors. In: Gamba P, Herold M (eds) Global mapping of human settlement—experiences, datasets, and prospects, pp 309–320. CRC Press, Boca Raton
- Dell’Acqua F, Iannelli G, Pittore M, Wieland M (2012) Inventory data capture tools—remote sensing (Draft V0.3), report published by Global Earthquake Model Foundation. Available at: http://www.nexus.globalquakemodel.org/gem-idct/files/wp-1-inventory_toolkit_development/userguide-remotesensing-v03.pdf. Accessed 15 Jan 2012
- Erdik M, Sesetyan K, Demircioglu M, Hancilar U, Zulfikar C, Cakti E, Kamer Y, Yenidogan C, Tuzun C, Cagnan Z, Harmandar E (2010) Rapid earthquake hazard and loss assessment for Euro-Mediterranean region. *Acta Geophys* 58(5):855–892
- Gamba P, Herold M (2009) Global mapping of human settlement—experiences, datasets, and prospects. CRC Press, Boca Raton
- Gamba P, Dell’Acqua F, Lisini G (2009) BREC: the built-up area recognition tool. In: Proceedings of urban remote sensing event, 2009 Joint, pp 1–5. 20–22 May 2009. doi:10.1109/URS.2009.5137593
- Giuliani G, Peduzzi P (2011) The PREVIEW global risk data platform: a geoport to serve and share global data on risk to natural hazards. *Nat Hazards Earth Syst Sci* 11:53–66
- Glade T, Anderson MG, Crozier MJ (2005) *Landslide hazard and risk*. Wiley, New York
- GPW (2004) Center for International Earth Science Information Network (CIESIN), Columbia University; International Food Policy Research Institute (IFPRI); The World Bank; and Centro Internacional de Agricultura Tropical (CIAT) Global rural-urban mapping project, Version 1 (GRUMPv1): population count grid. Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University. Available at <http://sedac.ciesin.columbia.edu/gpw>. May 21, 2012
- Huyck C, Esquivias G, Gamba P, Hussain M, Odhiambo O, Jaiswal K, Chen B, Yetman G (2011) Survey of available input databases for GED, GED4GEM deliverable D2.2, GEM foundation. Available at <http://>

- www.nexus.globalquakemodel.org/ged4gem/posts/ged4gem-deliverable-d2.2-survey-of-available-input-databases-for-ged/
- Jaiswal KS, Wald DJ (2008) Creating a global building inventory for earthquake loss assessment and risk management. U.S. Geological Survey Open-File Report 2008-1160, 103 p
- Jaiswal KS, Wald DJ (2010) Development of a semi-empirical loss model within the USGS prompt assessment of global earthquakes for response (PAGER) System. In: Proceedings of the 9th US and 10th Canadian conference on earthquake engineering: reaching beyond borders, July 25–29, 2010, Toronto, Canada
- Jones BG, Lewis BD (1990) Estimating size distributions of building areas for natural hazards assessments. *Earthq Spectr* 6(3):497–505
- Jones BG, Nicolaides CN (1988) Buildings at risk in the Whittier Narrows, California earthquake. *Earthq Spectr* 4(1):35–42
- Jones BG, Manson DM, Mulford JE, Chain MA (1976) The estimation of building stocks and their characteristics in urban areas: an investigation of empirical regularities. Program in Urban and Regional Studies at Cornell University, Ithaca, NY
- Jones BG, Manson DM, Hotchkiss CM, Savonis MJ, Johnson KA (1987) Estimating building stocks and their characteristics. Institute for Social and Economic Research, Program in Urban and Regional Studies, Cornell University, Ithaca, NY
- Leach JD (2001) Prediction of building count and dimensions from U.S. Census data using multiple regression. M.S. Thesis, Department of Geography, Virginia Tech. University, 65 p
- Lirer A, Petrosino P, Alberico I (2010) Hazard and risk assessment in a complex multi-source volcanic area: the example of the Campania Region, Italy. *Bull Volcanol* 72(4):411–429
- Merz B, Thielen AH, Gocht M (2007) Flood risk mapping at the local scale: concepts and challenges. *Adv Nat Technol Hazards Res* 25(Section III):231–251
- Pesaresi M, Gerhardinger A, Kayitakire F (2008) A robust built-up area presence index by anisotropic rotation-invariant textural measure. *IEEE J Sel Top Appl Earth Obs Remote Sens* 1(3):180–192
- Polli D, Dell'Acqua F, Gamba P (2009) First steps towards a framework for earth observation (EO) based seismic vulnerability evaluation. *Environ Semeiot* 2(1):16–30
- Salvatore M, Pozzi F, Ataman E, Huddleston B, Bloise M, Balk D, Brickman M, Anderson B, Yetman G (2005) Mapping global urban and rural population distributions, estimates of future global population distribution to 2015. Food and Agriculture Organization of the United Nations, Environment And Natural Resources Working Paper n. 24
- Schmidt C (2001) CatNet: a worldwide natural hazard atlas on the internet, Esri International User Conference. Available on line at <http://proceedings.esri.com/library/userconf/proc01/professional/papers/pap928/p928.htm>
- Schneider P, Schauer B (2006) HAZUS—its development and its future. *Nat Hazards Rev* 7(40):5. [http://dx.doi.org/10.1061/\(ASCE\)1527-6988\(2006\)7:2\(40\)](http://dx.doi.org/10.1061/(ASCE)1527-6988(2006)7:2(40))
- Schneiderbauer S (2007) Risk and vulnerability to natural disasters—from broad view to focused perspective, Ph.D. Thesis, Free University of Berlin
- Wald DJ, Worden BC, Quitoriano V, and Pankow KL (2005) ShakeMap manual: technical manual, user's guide, and software guide: U.S. Geological Survey, Techniques and Methods TM12-A1, p 132 (<http://pubs.usgs.gov/tm/2005/12A01/pdf/508TM12-A1.pdf>)
- Waters D, Cechet B, Arthur C (2010) Role of exposure in projected residential building cyclone risk for the Australian region. *IOP Conf Ser Earth Environ Sci* 11. doi:10.1088/1755-1315/11/1/012022
- Wood N (2009) Tsunami exposure estimation with land-cover data: Oregon and the Cascadia subduction zone. *Appl Geogr* 29(2):158–170