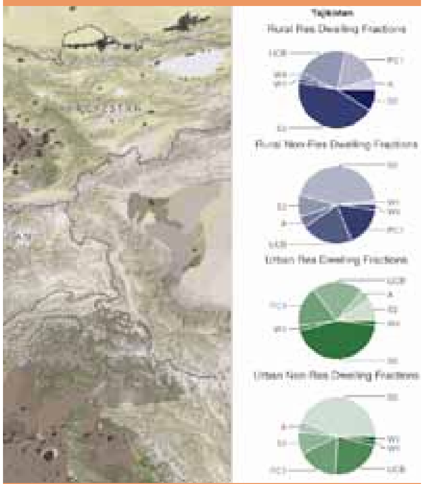
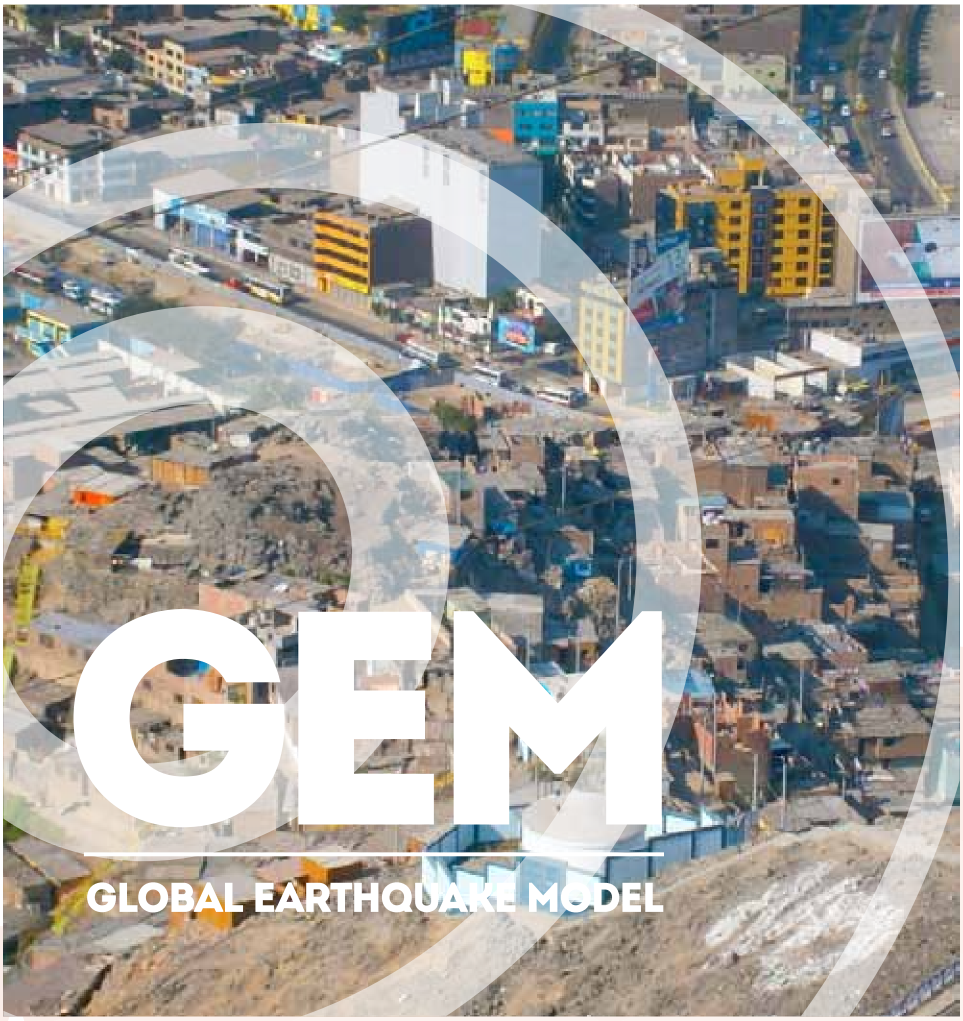
GEM TECHNICAL REPORT
2014/10 V1.0.0

Global Exposure Database – Scientific Features

P. Gamba



EXPOSURE
MODELLING



Global Exposure Database: Scientific Features

Technical Report 2014-10

Version: 1.0

Date: December 2014

Author(s)*: P. Gamba

(*) Authors' affiliations:

Paolo Gamba, University of Pavia

Rights and permissions

Citation advice

Gamba P. (2014), Global Exposure Database: Scientific Features, GEM Technical Report 2014-10 V1.0.0, 46 pp., GEM Foundation, Pavia, Italy, doi: 10.13117/GEM.EXP-MOD.TR2014.10.

Photo credits

Large cover image: © Nicola Tarque R.

http://www.globalquakemodel.org/

ACKNOWLEDGEMENTS

The authors of this report are grateful to all GED4GEM consortium partners for their contributions to the project and its deliverables, summarised in this report: Center for International Earth Science Information Network (CIESIN), Global Urban Observatory (GUO) of UN-HABITAT, ImageCat Inc., Joint Research Centre of the European Union (JRC), and advisors: US Geological Survey (USGS), EUCENTRE and Geoscience Australia (GA).

ABSTRACT

The aim of this report is to provide a comprehensive documentation of the GED4GEM project and its results, the Global Exposure Database (or GED). Although most (if not all) of the information provided in this document is available in the deliverables of the GED4GEM project (http://www.nexus.globalquakemodel.org/ged4gem/posts/ged4gem-deliverables/), this document offers a unique possibility to look at the various scientific aspects of the project. It also attempts to provide clear answers to the main questions that might be posed by potential users of the GED, specifically about the data sets used, the methodologies employed and the completeness of the database.

Keywords

exposure; GED4GEM; taxonomy; buildings; dwellings; inference.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

Page

1 Introduction to the Global Exposure Database built by the GED4GEM project

The aim of this document is to introduce the Global Exposure Database (GED for short) to interested users, and explain how it was designed and developed by the consortium of the project "A Global Exposure Database for GEM" (GED4GEM). This document serves to promote the use of the GED by explaining how it works, what it contains and how to interact with it, how to visualize and extract the data, and lastly how to propose amendments, improvements, corrections, and provide datasets to make the collection more accurate, complete and useful.

1.1 Why does the world need a GED?

The GED provides a spatial inventory of exposed assets for the purposes of CAT (Catastrophe) modelling and loss estimation. The GED allows for the assessment of risks and the mapping of hazards by interested parties such as non-profits, insurers, reinsurers, insurance brokers, countries, regional governments, and world financial institutions. These datasets are difficult to construct given the scarcity and inhomogeneity of base data available on the web or from government or commercial suppliers. When available, these data are geographically sparse, must be fused from disparate sources and frequently are not distributed with adequate metadata to assess data quality, initial source, or fitness of purpose. In 2010 the GEM Foundation issued a request for proposals to build a comprehensive repository of exposure data that included requirements to 1) document, and where possible, collect all relevant sources of information, 2) homogenise the data from sources to make them suitable and available for risk assessment tools such as the OpenQuake-engine. The GED4GEM consortium, led by the University of Pavia, was successful in proposing a system to be implemented within the GEM platform. This system was made available to GEM by the end of 2013, with the ability to be updated, improved and enhanced by a distributed effort, moving forward.

1.2 The GED4GEM team

The partners of the GED4GEM consortium comprised some of the major players in the definition, population and management of global databases, either for risk analysis or for other related research. They were the University of Pavia, leading the project, the Centre for International Earth Science Information Network (CIESIN), Columbia University, the Global Urban Observatory of UN-HABITAT, the Joint Research Centre of the European Union, and ImageCat Inc. Additionally, three consulting partners were also included as advisors, namely EUCENTRE, the European Centre for Training and Research in Earthquake Engineering, USGS, the United States Geological Service, and Geoscience Australia.

1.2.1 University of Pavia (UNIPV), Department of Electrical, Computer and Biomedical Engineering

The Remote Sensing Group of the Department of Electronics of the University of Pavia has been developing in the past years many techniques devoted to the analysis and interpretation of remote sensed data in urban areas. Its main activities are in the field of disaster management using earth observation data within and around urban areas, urban mapping and classification, joint three-dimensional and two-dimensional analysis of urban area, multi-sensor data fusion for urban characterization and vulnerability mapping. Particular emphasis is placed on reliability and speed of responses in crisis situation. In practical terms, this translates into reliable communication networks, sensor networks collecting distributed information, analysis of data acquired by airborne and spaceborne sensors, for vulnerability and damage assessment, or for environmental monitoring.

In the area of Remote Sensing, currently pursued activities are on semi-automated processing of satellite data and automated detection and gauging of 2- and 3-dimensional characteristics of buildings, bridges, roads and other man-made features. The Group has also started developing algorithms for automated estimation of damage level caused by natural disasters exploiting data from very high resolution radar and optical satellite sensors; more recently, methods for detection and tracking of targets in airborne radar sequences have been developed.

1.2.2 Center for International Earth Science Information Network (CIESIN), Columbia University

CIESIN is a research centre within the Earth Institute of Columbia University, based at the Lamont campus in Palisades, New York. CIESIN serves as the World Data Center for Human Interactions in the Environment of the International Council for Science (ICSU) and operates the NASA Socioeconomic Data and Applications Center (SEDAC). CIESIN scientists conduct interdisciplinary research at the intersection of the natural and social sciences, addressing critical problem areas such as multi-hazard vulnerability assessment, climate change impacts, environment and security, health-environment interactions, land use/land cover change studies, and Earth and space science informatics. CIESIN co-manages the Data Distribution Center of the Intergovernmental Panel on Climate Change (IPCC) with partner organizations in the UK and Germany, and hosts the Population-Environment Research Network (PERN) in collaboration with the International Union for the Scientific Study of Population (IUSSP).

CIESIN flagship data products include the Gridded Population of the World (GPW) dataset, used widely in both research and applications to understand population distribution at regional and global scales, and the Global Rural Urban Mapping Project (GRUMP) data collection, which provides a unique view of the spatial patterns of urban areas and other human settlements around the world. CIESIN is co-leading development of an open access global roads data collection, as part of a working group of the ICSU Committee on Data for Science and Technology (CODATA). CIESIN was a major contributor to the Global Natural Disaster Risk Hotspots study of the World Bank and the ProVention Consortium published in 2005 and has since participated in other major interdisciplinary disaster risk assessments. CIESIN is also active in numerous international initiatives on such topics as environmental indicators, poverty mapping, soils information, slum mapping, biodiversity informatics, international environmental agreements, peace and conflict data, climate change adaptation, polar data management, and open data access and data stewardship.

CIESIN is active in the development of the Global Earth Observing System of Systems (GEOSS), including participation in the Architecture Implementation Pilot (AIP), the Data Sharing Task Force, and several other Group on Earth Observations (GEO) tasks. CIESIN is a founding member of the Global Spatial Data Infrastructure Association, a university member of the Open Geospatial Consortium, and a supporting organization of the United Nations Geographic Information Working Group.

1.2.3 Global Urban Observatory (GUO) of UN-HABITAT

The Global Urban Observatory coordinates the Monitoring of the Habitat Agenda and the Millennium Development Goals as well as the implementation of the Agency Strategic and Institutional Plans such as the Medium-Term Strategic and Institutional Plan (MTSIP) for 2008-2013. GUO coordinates all activities pertaining to the production of reliable and up-to-date urban indicators at regional, country and city levels, including slum settlements that address the challenges of rapid urbanization and human settlements issues in the world.

GUO carries out activities pertaining to the Monitoring of the Habitat Agenda and the Millennium Development Goals Target 11. It updates regularly the urban indicators programme (UIP) at the city level as well as the slum indicators at country level. GUO coordinates use of UIP and MUIP information for the preparation of: the State of the World Cities Report (SWCRs); the Global Report of Human Settlements (GRHSs); the Global Water and

Sanitation Reports; and the United Nations Secretariat General reports on the MDGs. It ensures that reliable and up-to-date urban information are available for reports, messages and speeches of the Executive Director and other Representatives of the Agency. It represents the Agency at inter-agency meetings, including the MDGs Inter-Agency Expert Group (IAEG), the UN Statistical Commission (UNSC), the Common Coordination of Statistical Activities (CCSA), ECOSOC, CSD and others.

1.2.4 Joint Research Centre of the European Union (JRC), ISFEREA Action

As a service of the European Commission, the JRC is a reference centre of science and technology for the European Union. Its general mission is to provide customer-driven scientific and technical support for the conception, development and monitoring of EU policies. Its Institute for the Security and Protection of the Citizen (IPSC) contributes to the development of European Security Research and provides scientific and technologic support to EU security policies. The IPSC has also specialized in addressing the crisis management cycle of international disasters in support of the External Relation services of the Commission. The crisis management activities address all the phases of the crisis cycle including early warning, emergency response, damage assessment, support for reconstruction and most recently disaster risk assessment.

Isferea is the name of a JRC team of researchers focusing on geo-spatial and image analysis and novel field data collection methods. The research focuses on the detection and mapping of the constructed environment (or built up) and its characterization. Isferea's main activity focuses on optical satellite and aerial photography as well as last generation of SAR imagery. In addition to image analysis, Isferea has also assembled tool kits of hand held devices for field data collection such as GPS, digital cameras, video cameras. Isferea has then developed procedures to integrate the field data with the complementary information extracted from satellite imagery.

Most of Isferea application areas relate to the analysis of settlements and the different phases of the crisis management cycle. Rapid alerting, emergency response and damage assessment have become operational tasks for Isferea while disaster risk is a novel and challenging new area to be covered within the crisis management cycle. Within the disaster risk equation, Isferea has addressed the mapping of "physical exposure" that is the constructed environment. Two main strategies are used for this task. First, map the built up area using a procedure that is applicable globally and that produces consistent results globally. The second strategy aims to assess the building stock using procedures that are typically location dependent.

1.2.5 ImageCat, Inc.

ImageCat, Inc., an advanced technology company, specializes in innovative solutions to risk assessment and management. Established in March 2000, the company has already developed a reputation for innovative solutions in earthquake risk assessment and reduction, quantification of the built environment, post-disaster damage assessment, and transportation security.

An innovative flagship for the company is the application of remote sensing technologies to risk assessment and risk reduction. Supported by grants from the National Science Foundation, ImageCat's cutting-edge research focuses on the design of tools to better quantify the built environment and its exposure to risks associated with earthquakes, hurricanes, and other disasters. Current research focuses on the application of high-resolution remote sensing data to quantify the shapes and heights of residential, commercial, and industrial structures.

A leader in integrating remotely sensed data into Geographic Information Systems (GIS), ImageCat technology "fuses" visually accurate satellite and other data with relevant geo-referenced material to provide quantitative information on numbers of structures, proportion of residential, commercial, and industrial development, numbers of households, income levels, etc. Our innovative strategies for melding various data sets offer clients the most comprehensive data package available on the structural and human assets in a geographic area.

2 What can a user obtain from GED4GEM?

The GED provides information about two main assets at risk: residential population and residential buildings. Potentially, it can also include information about non-residential population and buildings, although the amount of public information for these two additional assets is currently quite limited.

The data included in the GED is the result of an extensive effort by GED4GEM partners in 1) analysing existing datasets, 2) selecting those that were most suited to the GED, 3) implementing population strategies to extrapolate the necessary attributes in format suitable for characterizing building and population exposure, and, in tandem, 4) homogenizing the data from various sources such as GEM Regional Programmes and GEM partners into a consistent format.

In general, the GED is divided into four different levels, which are populated from different data sources and using different techniques. Each level has a different geographical scale as for the statistical consistency of the data it contains, and has a different level of completeness. Each level is thus appropriate for a different use:

- Level 0 – a representation of the population and buildings on a 30" by 30" grid, with information about the buildings coming from statistics available at the country level. The building distribution is thus the same for each element of the grid belonging to a given country, with a binary difference between "rural" and "urban" area. The actual building counts (for each building typology) in each element of the grid are different, because the same distribution is modulated according to the residential population number available for that grid element.
- Level 1 – a representation of population and buildings on a 30" by 30" grid with information about the buildings available using the sub-national statistics (e.g., for regions, states, provinces or municipalities according to the different countries). Building distributions at this level are consistent at the sub-national scale, and, again, they are different for rural versus urban areas.
- Level 2 – a representation where each element of the same 30" by 30" grid includes enough information to be consistent by itself, and no distribution on a bigger geographical scale is used. This case corresponds to the situation when all building counts are actually obtained not by means of a disaggregation of a distribution available on a wider area on the elements of the grid but, on the contrary, by aggregating building level data, possibly available for the area of interest.
- Level 3 – a representation at the single building level, including all the possible information about each building, such as structural, occupancy and economic variables.

The four levels of the database are independent from each other. However, each level could be used to transfer information into other levels, if necessary.

As the GED levels include different information at different spatial scales, the data ingested into the GED come from different databases. Given the lack of standardization amongst the datasets available to ingest into a GED, algorithms able to process any given information source are simply not possible. However, specific population strategies for each level were analyzed and optimized within the GED4GEM project. The implementation of these strategies depended on the availability of the datasets to be used, the development of the IT infrastructure, and the actual engineering details of the database.

This is the reason why, in the following sections, there is a description of the actual structure of the database, together with a rough yet complete explanation on how the information about population and buildings are stored.

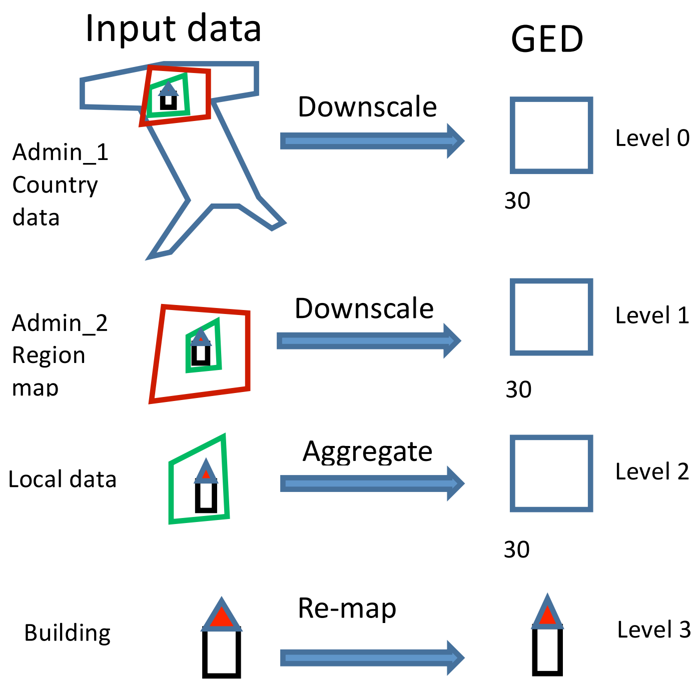


Figure 2.1 Conceptual overview of data entry and GED4GEM information content levels.

2.1 GED Level 0/1 Data Description

The levels 0 and 1 of GED are based, geographically speaking, on global grid of 30" x 30" resolution (nearly 1 x 1 km at the equator), stored in the database and represented by the grid cell's midpoint (called "grid point"), using its latitude and longitude. This grid includes only land areas; oceans, inland water bodies, and uninhabited places such as Antarctica are excluded. The exact geographic extent of the database is 180W, 180E, 85N, 58S. The grid geometry and attribute information is generated using the "Land/Geographic Unit Area Grids", "Population Count Grids; Year 2000", and "Urban Extents Grid" datasets from the SEDAC Global Rural-Urban Mapping Project, Version 1 (GRUMPv1, see section 3.2.1).

At level 0 population counts on the grid, a (rough) subdivision between residential, non-residential and transit population along the day, dwelling or building fractions at the country level, as well as statistical information about the average number of people per dwelling, the average number of dwellings per building (for each building typology) and the average size of a dwelling or the average floor per capita, are stored in the database. At level 1, the same database structure is replicated, but with dwelling or building fractions referring to sub-national administrative units. Level 0 and level 1 therefore require that the GED store information about national and sub-national administrative units, population counts, population ratios, dwelling and/or building fractions, national statistical facts about dwelling size and occupation as mentioned above. By means of a combination of the abovementioned information, building counts can be obtained at the grid level. These basic elements of the database are described in the next sections. The datasets used to extract this information are instead described in Section 3.

2.1.1 National and Sub-National Administrative Units

The geometry and attribute information of countries and their administrative units down to the finest available level are all stored in the GED. The parent-child relationship between these administrative units is also stored. Any information is time stamped and the database has the ability to store changes in boundaries and names over time.

In addition to political and administrative region, a construct named “geographic region” is defined in the database so that specific information about buildings/dwelling and population can be applied to a specific geographic area. The total population/building counts for that geographic region, computed by aggregating the population/building counts for all the grid cells within that region, can be also stored.

2.1.2 Population and population ratios allocations

The GED Level 0 and 1 store the population on the abovementioned grid as the total people count in that grid. GED is not responsible for the selection for these numbers, which are extracted from existing global databases, as described in section 3. Additionally, as important information for risk management purposes, the GED stores the day time, night time and transit population ratios, at a specific geographic region level, for each combination of urban/rural categorization and occupancy classes. Since there are only two occupancy classes, residential and non-residential, for each geographic region, there will be four sets of day/night/transit population ratios: urban/rural vs. residential/non-residential.

2.1.3 Dwelling and Building Fractions

The GED usually stores dwelling fractions for a level 0 or level 1 area. These fractions are obtained using different methodologies as described later on in this report, and are exploited to estimate building fractions for different building types by using statistical facts about a country, such as the average number of dwellings per building typology. Please note that the GED database only stores the distribution of dwelling fractions but not the procedure or original data set(s) from which it is derived.

Alternatively, building fractions may be stored that can be used directly to distribute the total building count among different building typologies.

A set of dwelling/building fractions, in a specific study region and for a specific urban/rural and occupancy class combination (e.g. urban/residential, urban/non-residential, rural/residential, or rural/non-residential), is defined in the GED database as a distribution group. Several fields to describe a distribution group, such as building fraction, dwelling fraction, and the data sources from which the attributes are derived, date of the data source, and the units of different fields are stored. Multiple distribution groups (i.e. different set of dwelling factions) can be stored in the GED database. Additional information and facts about a mapping scheme, such as total number of dwellings, total number of buildings etc., and their sources are also stored.

2.1.4 Statistical information at the country level

The GED stores statistical information at the country level, with the possibility to include more geographically refined information in the future. Specifically, the information stored is: a) the average number of people per dwelling; b) the average number of dwellings per building according to building typologies; c) the average size of a dwelling; the average floor per capita. Each of these numbers, wherever possible, is included for urban/rural area distinctly, and for residential/non residential occupancy types. Additionally, and for exposure computation purposes, the GED includes information about the reconstruction cost per square meter at the national level, ether for each of the building typologies or, more frequently, an average value valid for all of them.

2.1.5 Population Allocation and Building Count by Building Type at Grid Cell Level

The population allocation by building type and computation of building counts by building type at grid cell level within a study region can be computed using the abovementioned information stored into the GED. Note that, differently from the data mentioned in the previous sections, these values are not explicitly stored in the database for level 0 and level 1, but can be easily computed. Specifically, for each grid cell in a study region, the day/night/transit population allocation ratios for each building type can be computed by multiplying the fraction of that building type by the day/night/transit population ratios of that grid cell (which can be obtained from the appropriate set of population allocation ratios at the corresponding geographic region level for the given urban/rural and occupancy category of the grid cell). Using these ratios and the population count of the grid cell, the day/night/transit population counts for each building type within a grid cell can be computed. Based on that, the building counts by building type can be computed for each grid cell in that study region.

2.2 GED Level 2 Data Description

The GED Level 2 data is stored using a database model based on and compatible with the OpenQuake-engine database schema used to store exposure data. This ensures that the information and Level 2 can be directly and easily used as input to the OpenQuake-engine.

Therefore, at the Level 2 each element of the grid includes information about the building count for each of the building typologies in that specific location. The building distribution can be different for each grid element, or change irrespectively of any administrative/political boundary. Moreover, building counts are not obtained from population counts, and therefore no statistical information at the national or sub-national level is needed. These average values are instead substituted by the same information at the local (grid element) level.

Accordingly, the level 2 data stores, for each grid element, a record including building counts, the taxonomy used to describe the corresponding building types, the category of data provided (population or building) and the type and units of area used to store the information. Additionally, occupancy is included, to describe the occupants of a given asset over a specified period such as “day” or “night”, and one replacement cost value for each cost type can also be stored, if available. As for level0/1, each data is time stamped.

2.3 GED Level 3 Data Model Description

Level 3 data consist of records of structural characteristic of constructions, which can be buildings or other type of construction as long as they can be described using the GEM Building Taxonomy. “Object” was chosen instead of “Building” because it is a more accurate representation of this concept. Each record stores detailed structural information about an object using the GEM Building Taxonomy V2.0 (Brzev et al., 2013), with multiple "GEM Taxonomy..." fields storing the appropriate code representing corresponding structural characteristic.

In addition to structural characteristics, each object is associated with parameters like number of occupants at different time of day, cost of reconstruction, etc. These will allow calculation of economic and human losses for an earthquake event.

Level 3 data are expected to come from existing building databases or as a result of a ground survey. During a data collection deployment, often engineers take photos or videos while inspecting a site. These media files are stored on disk, but one or more records with metadata such as filename and type may be associated with the inspected object.

3 Sources used to populate the GED

The datasets used to populate the GED come from a huge variety of different sources, each one treated separately to extract as much as possible for the sake of the GED4GEM project. In the following we offer a brief introduction to the data sources used during the GED4GEM project, more details are available from the project deliverables.

Information about geographical entities

Geographical entities in the GED are extracted using existing global databases with political and administrative boundaries. By far the most important, and the only one available at the global level, is the GADM database.

3.1.1 Database of Global Administrative Areas (GADM)

GADM is a global administrative boundaries database created to support various geo-referencing uses and census data mapping. The collaborative group overseeing the creation and development of the GADM database strives to map the most up-to-date country boundaries at all administrative boundary level, national to sub-regional, for all nations and is open to the public for non-commercial use. The country names adhere to politically accepted boundaries and do not address or debate country name or extent for disputed regions.

Source: Database of Global Administrative Areas

Source Website: http://www.gadm.org/

Vintage: 2012

Key Metrics: Global country and administrative boundaries as polygons and raster grids.

Scale: Vector layers vary by country (Resolution: 30” arc-second)

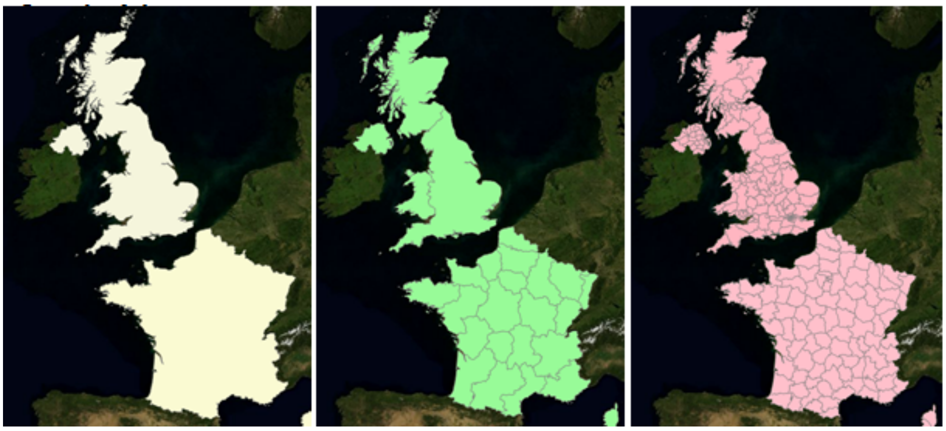


Figure 3.1 From left to right: Administrative boundaries level 0, level 1, and level 2 of the UK and France.

3.2 Rural and urban areas discrimination

In order to discriminate between urban and rural areas, databases obtained by using remotely sensed data have been considered. These data sets have several advantages. Whereas vector datasets as seen above are largely dependent on data collected from the ground, remote sensing databases are collected from the air and thus are less susceptible to bias. In the GED, the discrimination between urban and rural data is obtained by considering the human settlement extents extracted at the global level by the GRUMP project.

3.2.1 Global Rural-Urban Mapping Project-Urban Extent

The Global Rural-Urban Mapping Project (GRUMP) Urban Extent layer is a moderate resolution binary (presence/absence) map delineates the global rural/urban extent. The Urban Extent layer utilizes NOAA's Nighttime Light (NTL) 1994-1995 data to detect stable human settlement and ESRI's Digital Chart of the World's Populated Places (DCW) version 3 at the 1:1,000,000 scale for initial settlement points [CIESIN, *Urban Extent,* 2004]. For areas of inadequate to low electrical power sources the urban extents were extrapolated using a population-area ratio; Tactical Pilotage Charts were incorporated for urban delineation in Africa and Latin America for such reasons.

Source: Columbia University Center for International Earth Science Information Network (CIESIN)

Source Website: http://sedac.ciesin.columbia.edu/gpw/global.jsp

Vintage: 2000

Key Metrics: Urban and rural delineation in a binary raster.

Resolution: 30" arc-seconds

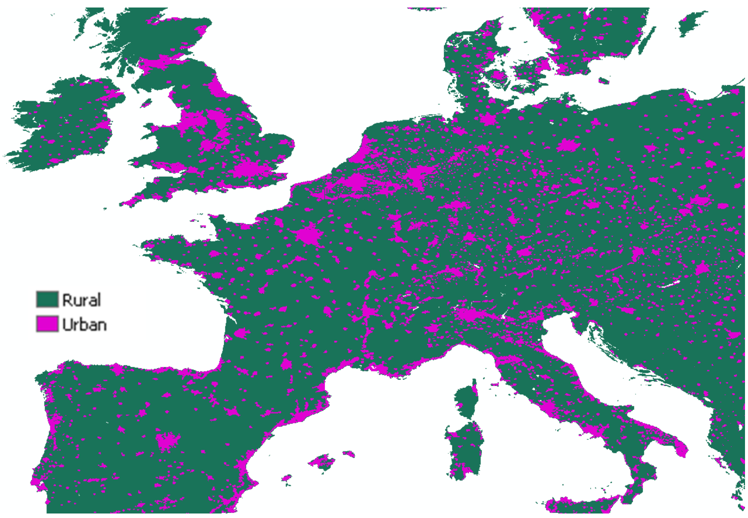


Figure 3.2 The delimited urban and rural extent over Europe.

3.3 Population data

Given the lack of building data at the global level, the GED uses population data as the basis, and by combining them with vector demographic data and statistical information at the national/sub-national level implements an efficient and effective approach to estimate the building counts. To this aim, the GED exploits raster data sets about population from the abovementioned GRUMP project but also from the GPW project, as described in the following.

3.3.1 Gridded Population of the World (GPW)

The GPWv3 demonstrates a spatially disaggregated distribution of human populations across the globe that is compatible with datasets from social, economic, and earth science fields. The output is unique in that the distribution of human population is converted from national or subnational spatial units (usually administrative units) of varying resolutions to a series of geo-referenced quadrilateral grids at a resolution of 2.5 arc minutes. GPW Future Estimates (GPWfe) contains estimated population grids for 2005, 2010, and 2015 based on simple population growth rates [CIESIN, 2005].

Source: Columbia University Center for International Earth Science Information Network (CIESIN)

Source Website: http://sedac.ciesin.columbia.edu/gpw/global.jsp

Vintage: Version 3, 2005, includes population estimates for 1990, 1995, 2000 and GPWfe provides estimates for 2005, 2010, and 2015.

Key Metrics: Even distribution of global population as a GRID.

Resolution: 2.5' arc-minute

Figure 3.3 GPW population density over Europe.

3.3.2 Global Rural-Urban Mapping Project-Population

The GRUMP global population distribution was developed to provide a global population estimate minimizing the impact of administrative boundaries. Therefore, GRUMP Population Grid improves GPW results by incorporating additional datasets which yield higher population concentrations surrounding observable developed areas. The additional datasets includes census administrative units and CIESIN's Urban Extent Mask developed from NOAA's Nighttime Light satellite images and the Digital Chart of the World's Populated Places (DCW) [CIESIN, *Population Grid,* 2004]. Using an interpolation with urban areas as a weighting function populations were reallocated from rural areas to known urban areas.

Source: Columbia University Center for International Earth Science Information Network (CIESIN)

Source Website: http://sedac.ciesin.columbia.edu/gpw/global.jsp

Vintage: 2000 (a 2010 estimate is under development)

Key Metrics: Population as a GRID

Resolution: 30" arc-seconds.

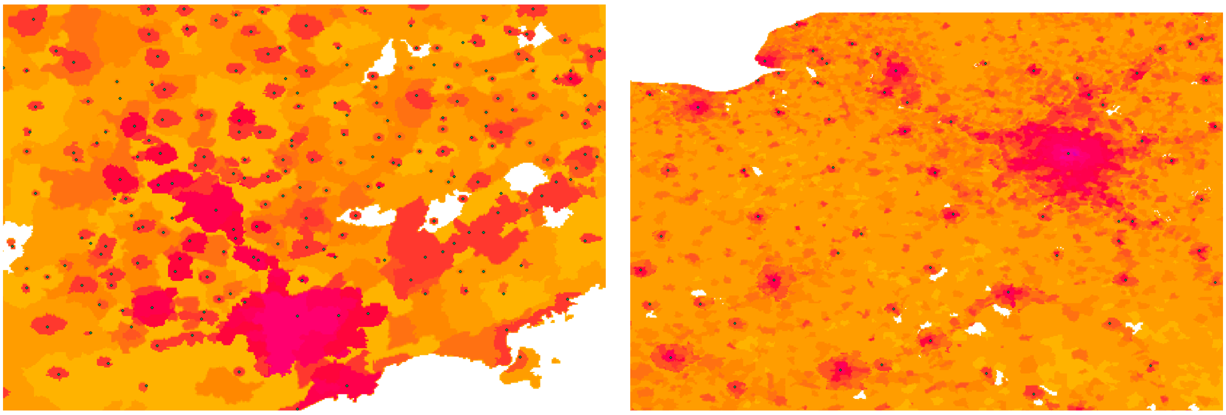


Figure 3.4 Examples of GRUMP 30" arc-second displays.

3.4 Demographic data

Demographic data is commonly attribute data associated with vector data described above, but can also be single numbers supplied on a national basis. Given the absence of global building exposure databases, demographic data is essential for estimating regional exposure. For example, if the number of buildings is unknown in a given country, population estimates by regional area or postal code are likely to still exist. Combined with such statistics as number of people per household and number of households per building, it will be possible to estimate the number of residential structures. Also, there may be statistics that characterize residential square footage per person. Given these figures, additional square footage would need to be added to characterize public, commercial and industrial use. Although population is easily found in datasets such as those discussed below, statistics such as average number of households per building or square footage per person are very difficult to locate.

3.4.1 Demographic and Health Surveys Data

Demographic and Health Surveys (DHS) are nationally-representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in the areas of population, health and nutrition. The indicators are presented in terms of national level statistics and for population subgroups such as those defined by age, education, marital status, economic status, urban/rural residence and region of the country. Household characteristics assessed include the household composition (how many people per household and their ages), educational attainment of household members, and school attendance ratios. Data on housing characteristics are also collected, including availability of electricity, water and sanitation facilities, as well type of flooring material and cooking fuel. Surveys also assess ownership of various durable goods, such as radio, television, refrigerator, bicycle, car and telephone. The respondent characteristics section provides basic demographic information about the survey respondents, including age, marital status, region of residence, level of education, religion and ethnicity. Educational attainment and literacy of respondents, access to mass media, employment status and occupation are also provided.

Source: Macro International

Source Website: http://www.measuredhs.com

Vintage: On-going, since 1984

Key Metrics: Household population, housing and respondent characteristics.

3.4.2 *Multiple Indicator Cluster Surveys Data*

Multiple Indicator Cluster Survey (MICS) is an international household survey programme developed by UNICEF. MICS data are collected during face-to-face interviews in nationally representative samples of households, generating one of the world's largest sources of statistical information on children and women. Since the mid-1990s, MICS has enabled more than 100 countries to produce statistically sound and internationally comparable estimates of a range of indicators in the areas of health, education, child protection and HIV/AIDS. MICS provides data at the national level that can also be disaggregated by various geographical, social and demographic characteristics.

Source: UNICEF

Source Website: http://www.childinfo.org

Vintage: On going, since 1995

Key Metrics: Indicators on access to water and sanitation, health, education, child protection and HIV/AIDS.

3.4.3 *Population and Housing Census Data*

A population census has quite a few intangible positive values. It is an opportunity for mobilizing the whole country and reaching even the most remote corners of it. In the life of many citizens a regular census is often the only time that the state reaches to them and asks them a question. Population censuses are important sources of information on population and housing, its age structure and its geographical distribution. Census data can provide denominators for indicators. The United Nations Statistical Division has been coordinating the preparation of the updated version of the *United Nations Principles and Recommendations for Population and Housing Censuses* as part of the 2010 World Programme on Population and Housing Censuses. Censuses, as a tool used by governments, have been around for few millennia and their essence, enumerating people and households, is one of the few activities that remained unchanged over all these centuries. The first set of *United Nations Principles and Recommendations for National Population Censuses (1958)* defines the census as the total process of collecting, compiling and publishing demographic data. During the 2000 round (2000-2005) of censuses, over 190 countries conducted a population census and an overwhelming majority intends to repeat it in the 2010 round (2010-2015) of censuses.

Source: Minnesota Population Center, University of Minnesota

Source Website: https://international.ipums.org/international-action/samples

Vintage: On-going, since 1960

Key Metrics: Population and housing data.

3.4.4 *UN-HABITAT Data*

UN-HABITAT's Global Urban Observatory (GUO) monitors global progress in implementing the Habitat Agenda and the global urban conditions and trends through the Urban Indicators Programme (UIP) that addresses the five chapters of the Habitat Agenda with a specific focus on shelter, social development and eradication of poverty, environmental management, economic development and governance. The programme addresses the urgent need to improve the world-wide base of urban knowledge by supporting Governments, local authorities and organizations of the civil society develop and apply policy-oriented urban indicators, statistics and other urban information. The programme also coordinates the monitoring of the Millennium Development Goals and activities pertaining to the production of reliable and up-to-date urban indicators at regional, country and city levels, including slum settlements that address the challenges of rapid urbanization and human settlements issues in the world. The GUO assists the Agency in the effort to become, by 2013 a premier reference centre for data collection, analysis, monitoring and reporting on sustainable urbanization.

3.4.5 United Nations Statistics Division Data

The Statistics Division provides a global centre for data on international trade, national accounts, energy, industry, environment and demographic and social statistics gathered from national and international sources. The Division regularly publishes data updates, including the Statistical Yearbook and World Statistics Pocketbook, and books and reports on statistics and statistical methods.

Source: United Nations Statistics Division

Source Website: http://unstats.un.org/unsd/default.htm

Vintage: On-going, since 1948

Key Metrics: Population of capital cities and cities of 100,000 and more inhabitants.

3.5 Inference

Given that none of the data sources discussed above can be simply used directly to develop a global exposure database, assumptions are required to assemble a suitable product. The primary building data required for assessing vulnerability includes structural characteristics, building size, building height, and building age. Structural characteristics are not well characterized in any of the datasets discussed above. Thus, for default data, structural characteristics must be inferred from other characteristics. Additional research will be required to establish the extent to which inference algorithms might be based on the raster datasets discussed above. The sources below are the known data sources for inference algorithms.

3.5.1 EERI World Housing Encyclopedia (WHE)

The World Housing Encyclopedia is developed by a consortium of EERI members throughout the world. Structural characteristics are characterized on a nationwide basis specifically with earthquake vulnerability in mind.

Source: World Housing Encyclopedia

Source Website: http://www.world-housing.net/

Vintage: On-going, since 2004.

Key Metrics: Structural information of various global building types.

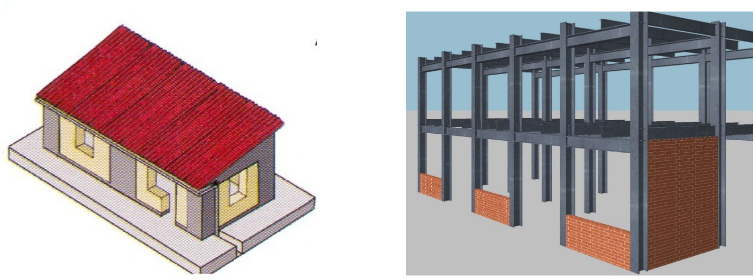


Figure 3.5 Reinforced Adobe (left) and Prefabricated Metal Construction (right).

3.5.2 USGS Prompt Assessment of Global Earthquakes for Response (PAGER)

USGS has expanded the general framework of the US risk assessment program HAZUS to the entire world. Occupancies are collapsed to simply urban and rural. Given that the objective of PAGER is casualty estimation, the inference schemes within PAGER represent an estimation of the percentage of population residing in any given structural type, not the percentage of buildings or square footage associated with a given structure type. The database draws on and harmonizes numerous sources: (1) UN statistics, (2) UN Habitat's demographic and health survey (DHS) database, (3) national housing censuses, (4) the World Housing Encyclopedia and (5) other literature (Jaiswal et al. 2010). The inference schemes from PAGER form a powerful basis for inferring structural types globally. The database is freely available for public use, subject to peer review, scrutiny, and open enhancement.

Source: USGS-Earthquake Hazards Program: PAGER

Source Website: http://earthquake.usgs.gov/earthquakes/pager/, http://pubs.usgs.gov/of/2008/1160/

Vintage: 2009

Key Metrics: Estimates fatalities and economic impacts immediately following large worldwide earthquakes.

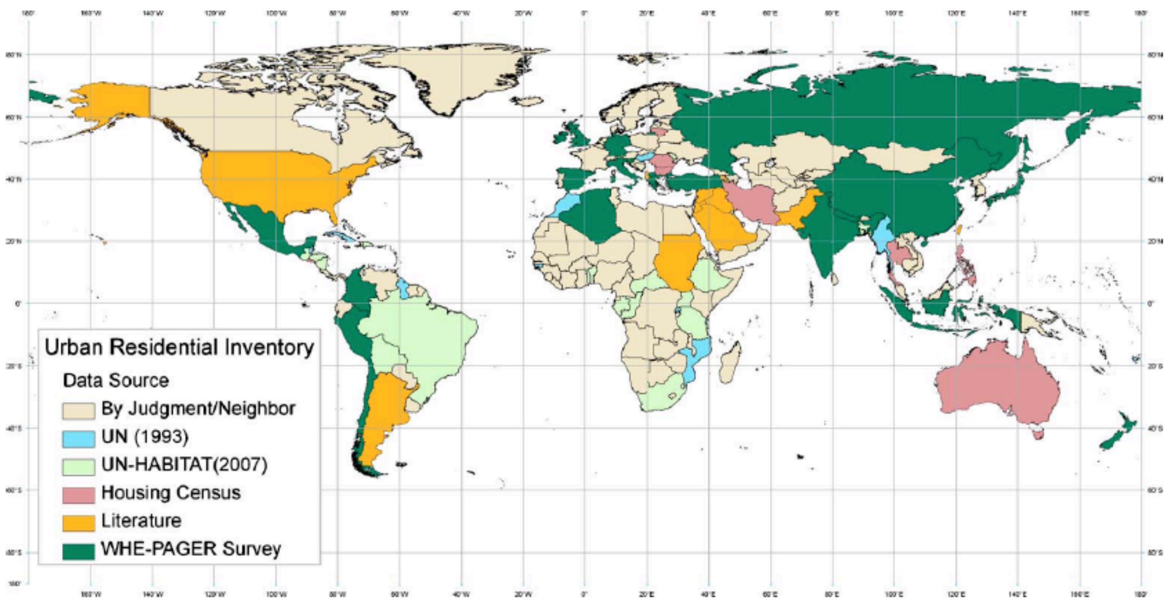


Figure 3.6 Urban residential inventory coverage by its source used within PAGER system for loss estimation.

3.6 Regional Data Sets

Regional datasets typically track both the natural and human-made environment in GIS or raster coverage such as land use, political jurisdictions, soils, and elevation. Certain national data sets such as bridge, pipeline, schools, or other essential facilities data may prove useful for national studies and will need to be integrated on a nation-by-nation basis. GED exploits mainly two of these regional datasets, provided by the NERA European project and the US-based HAZUS project.

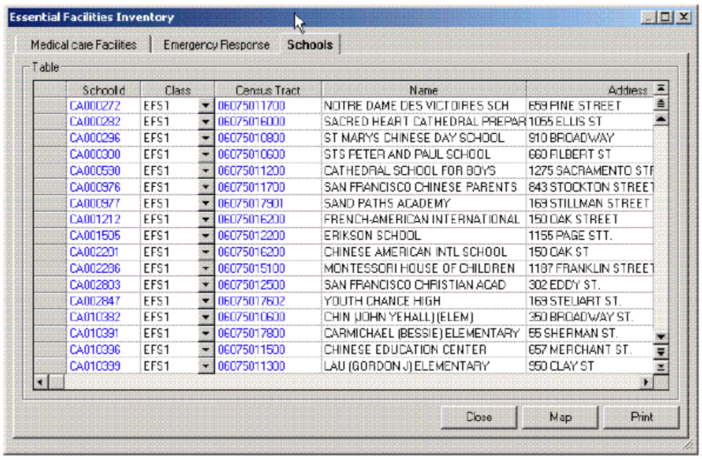
In this same category for the GED are included, although not yet available at the time of the end of the GED4GEM project, the databases obtained by the GEM Regional Programmes (e.g., EMME, EMCA).

3.6.1 NERA: Network of European Research Infrastructures for Earthquake Risk Assessment and Mitigation

NERA is a European FP7-funded project that has as one of its tasks that of developing a European building inventory database. Census data from each European country has been collected. The building attributes that are missing from these European data sources have been inferred from available attributes, mainly through expert judgment and elicitation (see work package 7 on www.nera-eu-org).

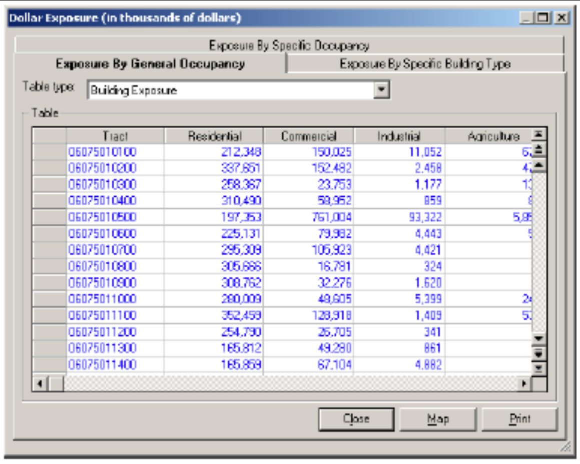
3.6.2 Multi-hazard Loss Estimation Methodology (HAZUS-MH)

The HAZUS-MH software (widely known as “HAZUS”) is designed to produce loss estimates for use by federal, state, regional and local governments in planning for multi-hazard risk mitigation, emergency preparedness, response and recovery. The two default inventory databases available within HAZUS are a) national listing of individual facilities located within US, e.g. dams, bridges, hazmat facilities etc., and b) general building stock datasets that are aggregated at county or census tract levels. The general building stock (GBS) includes residential, commercial, industrial, agricultural, religious, government, and educational buildings. The GBS dataset mainly consists of square footage by occupancy, replacement value by occupancy, building counts, and demographics. These default datasets are derived from publicly available data, eliminating fields of data that are not needed for the methodology. Many of the HAZUS default mapping relationships, e.g. occupancy to structure type as shown in next figure, are central to the derivation of the HAZUS GBS dataset.



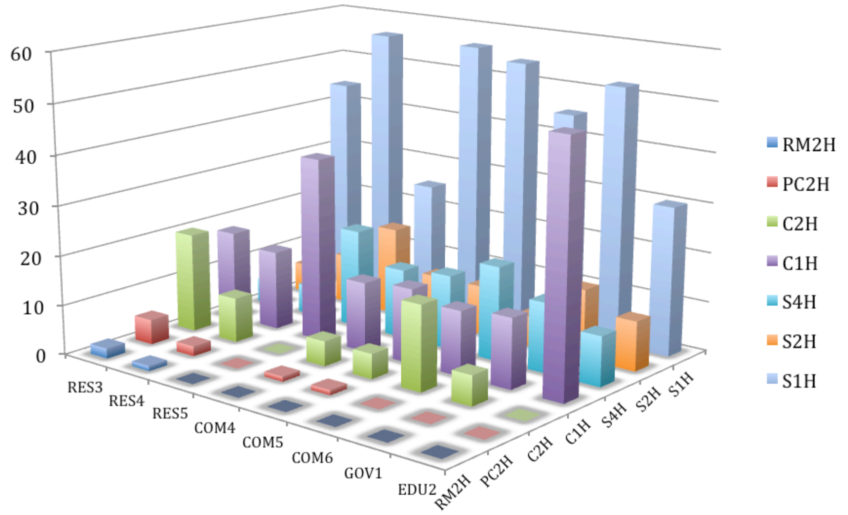
SchoolId	Class	Census Tract	Name	Address
CA000272	EFS1	06075011700	NOTRE DAME DES VICTOIRES SCH	659 PINE STREET
CA000282	EFS1	06075016000	SACRED HEART CATHEDRAL PREPAR	1055 ELLIS ST
CA000296	EFS1	06075010800	ST MARYS CHINESE DAY SCHOOL	910 BROADWAY
CA000300	EFS1	06075010600	STS PETER AND PAUL SCHOOL	660 FILBERT ST
CA000530	EFS1	06075011200	CATHEDRAL SCHOOL FOR BOYS	1275 SACRAMENTO STR
CA000976	EFS1	06075011700	SAN FRANCISCO CHINESE PARENTS	843 STOCKTON STREET
CA000977	EFS1	06075017901	SAND PATHS ACADEMY	169 STILLMAN STREET
CA001212	EFS1	06075016200	FRENCH-AMERICAN INTERNATIONAL	150 OAK STREET
CA001505	EFS1	06075012200	ERIKSON SCHOOL	1155 PAGE STT.
CA002201	EFS1	06075016200	CHINESE AMERICAN INTL SCHOOL	150 OAK ST
CA002236	EFS1	06075015100	MONTESSORI HOUSE OF CHILDREN	1187 FRANKLIN STREET
CA002803	EFS1	06075012500	SAN FRANCISCO CHRISTIAN ACAD	302 EDDY ST.
CA002847	EFS1	06075017602	YOUTH CHANCE HIGH	169 STEUART ST.
CA010382	EFS1	06075010600	CHIN (JOHN YEHALL) (ELEM)	350 BROADWAY ST.
CA010391	EFS1	06075017800	CARMICHAEL (BESSIE) ELEMENTARY	55 SHERMAN ST.
CA010396	EFS1	06075011500	CHINESE EDUCATION CENTER	657 MERCHANT ST.
CA010399	EFS1	06075011300	LAU (GORDON J) ELEMENTARY	950 CLAY ST

(a) Database of school facilities in USA



Tract	Residential	Commercial	Industrial	Agriculture
06075010100	212,348	150,025	11,052	6,
06075010200	337,651	152,482	2,458	4,
06075010300	258,367	23,753	1,177	1
06075010400	310,490	58,952	859	
06075010500	197,353	761,004	93,322	5,8
06075010600	225,131	79,982	4,443	5
06075010700	295,309	105,923	4,421	
06075010800	305,666	16,781	324	
06075010900	308,762	32,276	1,620	
06075011000	280,009	48,605	5,399	2
06075011100	352,459	128,918	1,409	5
06075011200	254,790	26,705	341	
06075011300	165,812	49,280	861	
06075011400	165,859	67,104	4,882	

(b) Exposure value of general building stock by census tract



(c) Sketch showing distribution of percentage of floor area for Model Building Types within Each Building Occupancy Class, High Rise, Post-1970, West Coast. The specific mapping scheme adopted here is after ATC-13 project.

Figure 3.7 Overview of HAZUS General Building Stock (GBS) inventory datasets and a mapping scheme taken from HAZUS-MH MR4 Technical Manual, 2010.

3.7 Local Datasets

In addition to the dataset discussed above, local data sets have been and can be ingested into the GED. Local datasets include city, sub-country or country-level datasets which track individual features in great detail. Examples include building point or footprint data, which may track the owner, area or number of occupants. In the US, for example, this data typically tracks square footage, date of construction, and the ISO building construction type. Many planning departments have embarked on efforts to render their cities in three dimensions, and local utilities typically have very detailed records that can be used to assess vulnerability. However, despite the prevalence of these types of sources, local municipalities can be quite resistant to release these sources, even if widely perceived as being for the common good. In addition, once these data are acquired, they typically require significant manipulation and interpretation by GIS experts, engineers, and local experts before they are suitable for analysis.

4 Methodologies employed to populate the GED

This section addresses the strategies employed to populate the different GED Levels. Users will be able, in addition to possibly replicate these approaches, to exploit the tools deployed by the IDCT (Inventory Data Capture Tools) consortium (see also section 6 and http://www.globalquakemodel.org/resources/use-and-share/tools-apps/).

4.1 Populating Level 0 data from statistical information

The basic GED level 0 data population strategy starts from exploiting statistical information about dwelling fractions at the country level, since these data are freely available in PAGER. However, the GED4GEM consortium has developed a strategy for the GED level 0 that does not limit itself to the information available in PAGER, which was developed for casualty estimation. GED4GEM improved it in two ways:

- by including information about dwelling fractions in countries where data are missing in PAGER and has been copied from neighbouring countries, by exploiting instead available statistical information from census, DHS (Demographic and Health Survey) and MICS (Multiple Indicator Cluster Survey);
- by computing and including building counts and fractions, starting from dwelling fractions, population counts and other ancillary available information.

The following sections provide an overview description of the methodologies currently considered for populating GED Level 0. The description does not include the technical details required for the fine implementation that are instead provided in the GED4GEM deliverables (http://www.nexus.globalquakemodel.org/ged4gem/posts/ged4gem-deliverables/).

Preliminary work consisted in searching available census, DHS and MICS data sets from the UN-Habitat data repository, cross-correlating with countries whose dwelling fractions are classified as “low” in PAGER ranking system. The priority is given to those countries where mapping schemes are obtained by geographical proxies, i.e. from dwelling fractions in one of the neighboring (or close enough) countries. Specifically, the analysis looked for the availability of information reporting on “wall typology”, “floor typology” and “roof typology”. In fact, a combination of these attributes, when available, can provide enough information to develop a proxy for structural characteristics of the surveyed dwellings. When not fully available, a subset of the same information might also be considered, as opposed to neglecting any possibility to extract dwelling fractions for that country.

4.1.1 Extract dwelling distributions from census and survey

The overall structure of this procedure is described by three processing steps:

1. extraction of statistically significant data from census, DHS and MICS records for a whole country;
2. mapping different combinations of wall/roof/floor materials into building classes;
3. application of the results of step 2 to the data after step 1 in an automated way and ingestion of the final dwelling fractions into the GED.

4.1.1.1 Extraction of significant statistical data from census, DHS and MICS surveys

Censuses and surveys such as DHS and MICS record information about dwellings as a means to map poverty and access to healthy sources of water, foods, and other basic human needs. All are designed to provide a statistically significant “snapshot” of the state of the population. Censuses are indeed complete, while DHS and MICS are samples. However, accurate statistics can be obtained, either by grouping the records (census) or weighting the samples in each class according to the population model used to design the survey (DHS and MICS). Furthermore, weights are computed (and applied) separately for rural and urban environments, enabling derivation of separate statistics.

Unfortunately, not all the information necessary for structural characterization is available, but a good number of proxies can be obtained. In addition to the already mentioned wall/floor/roof types, separation between residential and non-residential buildings is possible (although non-residential buildings are usually poorly sampled in the two data surveys). Similarly, the ownership of dwelling, year or period of construction, total dwellings, average number of people per dwelling and the total sample population are recorded, and sometimes the average area of the dwelling, the average number of rooms per dwelling and the average room size are also included. Finally, indexes like “overcrowding” can be extracted as a combination of these data.

4.1.1.2 *Mapping different combinations of wall/roof/floor materials into PAGER STR or GEM building taxonomy classes*

This step is based on a mapping from combinations of wall/roof/floor material to PAGER or GEM Building Taxonomy classes. A suitable strategy should be defined to cope with combinations that are not consistent, and for incomplete datasets (e.g., when only subsets of three basic material variables are available).

The overall strategy is based on tables assigning combinations to PAGER or GEM Building Taxonomy classes from all countries in a specific region. This approach reduces the burden to find a country-specific scheme that may or may not be available using inputs from technical literature or technical reports, and simplifies the problem by considering multiple information inputs from neighboring countries.

The issue of inconsistent or “impossible” combinations, i.e. combinations of floor and roof materials that are inconsistent with the way a house can be built, were treated on an ad hoc basis by looking for their statistical significance. If the percentage of dwellings was smaller than a given fraction, it was aggregated to other structurally significant classes. If, instead, the percentage was large, then further checks were performed with the authority that managed the corresponding survey.

4.1.1.3 *Final dwelling fractions extraction*

The final step of the procedure then fuses the information coming from the first and the second steps to provide a two column table with the relevant PAGER or GEM building classes for a specific country in the format that is suitable for fast inclusion into the GED.

4.1.2 Computing building distributions

The approach proposed for building count estimation is based on the availability of population counts on the GED 30” x 30” grid, as well as the dwelling fractions available in the original Level 0 data set, or obtained by means of the procedure discussed in the previous section.

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. 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. Once enough samples characterizing specific built environments are compiled, 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_{Zi} represent the total geographic area (km^2) covered by the *i*-th study zone Z_i and the term b_{Zi} 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 $b_R \approx E\left(b_{Z_i} \cdot \left(A_R / A_{Z_i}\right)\right)$, 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 human settlement maps extracted from remote sensing data can be used as a starting zonation model to apply such schemes.
- *Building counts from cadastres 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 into the GED4GEM database. Similarly, certain countries with national cadastres 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 use 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 $\mathrm{b(i,j)} = \mathrm{f}\big(\mathrm{p(i,j)}\big)$ may be estimated and directly applied only if enough building count samples are available for a specific geographical area (e.g., a country) to derive region-specific regression parameters.

4.2 Populating Level 1 data from statistical information

The procedures described in the previous section at the country level can be applied also to sub-national statistical information extracted from the same census, DHS and MICS surveys. It is expected that the dwelling fractions extracted at the country level according to those procedures can be refined at any sub-country level for which the data subset is still statistically significant.

As a matter of fact, as long as the subset of the original country records belonging to any specific sub-national administrative subdivision is statistically significant, the procedure in Section 4.1 can be applied to this subset instead of to the whole country set, and provide significant dwelling fractions, building fractions for that specific sub-national zone.

4.3 Populating Level 2 data from building-to-building databases

Populating GED using building-to-building databases requires strategies for both available datasets and those that still need to be created. From the GED4GEM point of view there are two types of existing building-to-building (B2B) databases: those that contain structural information on buildings (B2B_1) and those that do not (B2B_0). The latter are considered for populating Level 2; the former for populating Level 3, and are respectively described in the following section.

"B2B_0" building stock databases are generated from at least four types of authorities: Land parcel systems authorities (those managed by legal entities that are empowered to collect taxes), city and planning authorities, cartographic institutions, and increasingly international organization assisting low income countries to generate their building stock information layers. These datasets exist in abundance but are often not publicized or shared. Access to this information remains a challenge for a number of reasons. Land parcel systems (cadaster) are often prevented from releasing detailed data because of privacy issues. Other institutions produce the data under commercial contracts or licenses, or also face copyright or privacy constraints and may not make the data available unless proper institutional arrangements are set up. The new crowdsourcing community has the potential to generate massive B2B information that should be considered by GED.

When available, "B2B_0" will be considered for populating GED Level 2. The inference scheme would be available from local sources, from local expertise and know-how, or – only if all other options are not available - inherited from higher level GED4GEM (i.e., Level 1). The strategy to populate GED4 with B2B_0 is included the following steps:

- ***Obtain the database and, if not accessible, obtain the metadata***. All countries experiencing seismic risk – for their own benefit and for the benefit of the community – should develop metadata of available building databases. The data may not be made available but their metadata should be openly available through GEM. Potential users may not find data directly in GED but this gives them the option to contact the authority that maintains the data. The metadata records should contain all relevant parameters such as the source of information used to generate the building database, encoding technical specifications, source of thematic information, authority holding the database, and information on the accessibility (e.g., contact name and phone number).
- **Describe and benchmark**. The strategy would then include the description and benchmarking of the datasets whose qualities need to be evaluated before ingestion. This work should be conducted for all countries that experience seismic risk. The benchmarking could be conducted by research institutions within the country with guidelines provided by GEM. Procedures and processes for this task could be defined within GEM. For those countries that consider building location a privacy issue, data could be released in aggregated form. The benchmarking is a critical step. It will have to be conducted using precise guidelines for both the locational (geometric) and thematic accuracy. For example, it is envisioned that many building footprint databases captured from open source

databases including OpenStreetMap intended for GED Level 3 will not comply with the geometric geometric/geographical specifications and would be rather be more suitable for GED Level 2.

4.4 Populating Level 3 data from building by building databases

Out of the many B2B databases available only those that also contain structural information ("B2B_1") will be considered for populating Level 3. "B2B_1" building stock databases are generated from specific surveys aimed at collecting structural information on buildings. A number of high income countries include collection of information in their censuses that can be used for input into GED Level 3. A crowd sourcing community has also started to generate building location, building height, and also building typologies that may be considered for use in GED Level 3. Finally, international organizations with mandate to address disaster risk including Global Facility for Disaster Risk and Recover (GFDRR) of the World Bank are funding exposure data capture projects at the building footprint level for Pacific Islands and the Caribbean using satellite imagery and Laser data. Exposure database at the building footprint level may be accessible and available from other international institutions or from institutions that link directly with governmental bodies such as the United Nations Development Programme, Global Risk Identification Programme (UNDP/GRIP).

The procedure for populating B2B_1 into GED follows the schema used in the previous section. The first two steps: (1) Locate the B2B database and (2) Describe metadata, are identical, hence they are not repeated. The benchmarking (step 3) will instead use the strict guidelines defined within GED Level 3, including a check on the data geo-coding, typology and completeness.

4.5 Level 0 and 1 data ingestion

Technically, the ingestion into the GED of level 0/1 data is performed by running three types of scripts: 1) Data verification, 2) Data ingestion, and 3) GED internal scripts. The **data verification scripts** serve two purposes: a) to check the consistency and validity of the candidate exposure dataset; b) to extract the required data from the candidate exposure datasets and convert it into a "GED-ready" format. The **data ingestion scripts** are used to ingest the extracted variables into the GED. There will be a family of scripts for each GED level and for each candidate exposure dataset format. Finally, **GED Internal scripts** aim at processing records available in fields filled from external data and analyse/treat/aggregated them into values to be inserted into other GED fields.

The verification scripts are the most demanding to write. Each of them works according to the following overall logic: it reads the information from the relevant candidate exposure datasets. The data undergo a number of checks of consistency between the spatial units and the taxonomies of the table compared to the fields of GED. If the checks are negative and the input data are found to be consistent with the fields of GED, the data are converted in machine-readable format and ready for ingestion. The relevant information can thus be ingested in GED using ingestion scripts. If the verification script identifies errors of consistencies it writes an error. The dataset needs to be re-processed for the correction of errors and verified again.

The complexity of the verification scripts depends on the formatting and the consistency of the candidate exposure databases. In fact, there will be as many verification scripts as there are data input formats. Verification script complexity depends on 1) GED Levels, 2) degree of compliance of the input data set with GED structure, variables and taxonomy, 3) degree of formatting of the files storing the input data sets, and 4) (in)completeness of the input dataset (e.g., whether it requires to inherit variable from other GED tables).

5 Coverage of GED4GEM

Although the previous section include data sources and methodologies, these data are not available everywhere, and the approaches sometimes could not have been applied to specific datasets, because of different technical and practical issues. As a result, the final delivered GED (the so called GED v 1.0) has some data gaps, as well as issues that have not been considered or completely covered. Indeed, an analysis of the gaps in the GED datasets and structure was done many times during the four years of the GED4GEM project. As a result, a first set of gaps affects the current usefulness of the GED, because there are countries or regions where it does not contain enough (or enough detailed) data. These gaps are likely to be fixed in the future, because the procedures to include new data into the system are already available, and it is ***just*** a matter to collect these data. The second set of gaps is a more problematic one, because it deals with complex situations that are not (completely) considered in the GED. They require some development (if not research) to be solved.

5.1 Missing level 0/1 data

The GED4GEM partners have been working hard to reduce the number of countries missing total information at level 0/1. The partners mobilized their resources in their geographical and other regions of influence to gather additional datasets primarily from population and housing censuses and household surveys as well as obtaining information on countries for which dwelling/building fractions could have been retrieved. These efforts significantly reduced the number of countries missing total information at level 0/1 from 163 to 93 summarized in Table 5.1 as follows:

Table 5.1 List of regions missing total information at level 0/1 at the end of the project

Geographic region	Total
Africa	12
N America	2
Caribbean, C & S America	23
SE Asia	4
Rest of Asia	16
Europe	24
Oceania	6
Other	6

5.2 Countries with missing data about dwelling size

The GED4GEM partners continued to source for and regularly updated the global list of “country facts” parameters as an on-going activity for the life of the project as part of its core functions of monitoring global urban conditions and trends through the Urban Indicators Programme. The updates included such parameters as the average number of people per dwelling, the average area of dwelling (m^2), the average floor per capita (m^2) and any other useful data taken from processed census/DHS/MICS data and any other data sourced within and outside the UN. In particular, the combined effort of updating the global list of “country facts” parameters was extremely successful, and only 13 out of 247 GADM 2 countries have missing data about dwelling size or the average number of people per dwelling at the national or sub-national level at the end of the project, as shown in Table 5.2.

Table 5.2 Number of countries missing data about dwelling size at the end of the project.

Serial No.	GADM2 Country Name	Geographic Region
4	Algeria	Africa, North
243	Western Sahara	Africa, North
141	Mayotte	Africa, East
234	United States Minor Outlying Islands	America, North
184	Saint Barthelemy	Caribbean
32	British Indian Ocean Territory	Asia, South East
215	Taiwan, Province of China	Asia, East
210	Svalbard and Jan Mayen	Europe, North
71	Faroe Islands	Europe, North
83	Gibraltar	Europe, South
96	Holy See (Vatican City State)	Europe, South
175	Pitcairn	Other
205	South Georgia and the South Sandwich Islands	Other

Table 5.2 shows that except for Algeria in North Africa and the Vatican City state in South Europe, the rest of the countries missing data about dwelling size are island states and/or dependent territories where any form of data is usually hard to come by. Dwelling size data at the national and/or sub-national level are usually commonly obtained from such data sources as population and housing censuses, household surveys, population-based country reports and more increasingly from free internet sources.

5.3 Graphical representations of level 0/1 available datasets

Here we provide a graphical representation of the level 0/1 dwelling and/or building fractions that have been collected by the GED4GEM consortium and not currently available in any other freely accessible exposure database. They add to those available by the PAGER system, which have been included and redefined according to the GED schema.

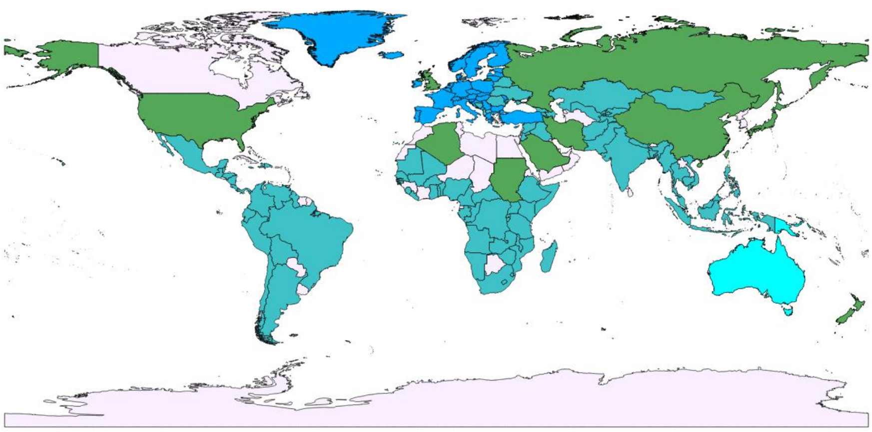


Figure 5.1 The geographical coverage of GED4GEM level 0 data (not considering PAGER data).

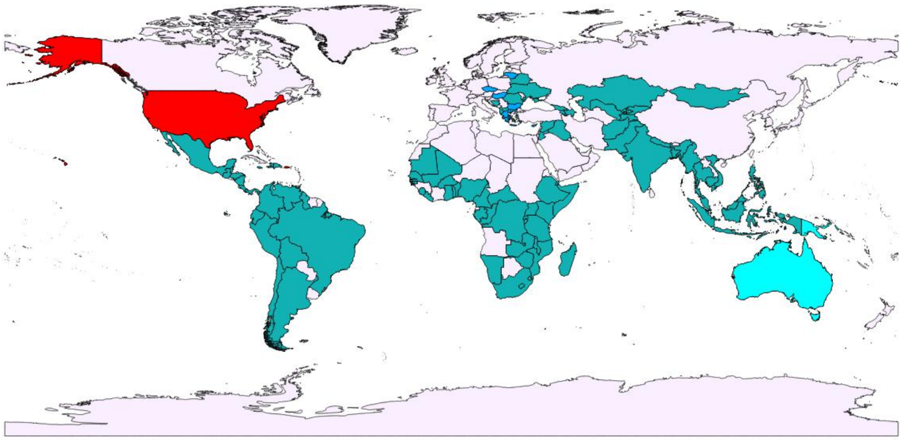


Figure 5.2 The geographical coverage of GED4GEM level 1 data.

In addition to the dwelling fractions, there are gaps with respect to the statistical information at the country level, namely:

- *Population ratio allocation*: the population allocation, which GED extracted from PAGER, is missing for a few countries (Kosovo, Spratly islands, Saint-Martin, South Sudan, Bonaire, Saint Eustatius and Saba, Sint Maarten, Saint-Barthélemy, Curaçao, Clipperton Island, Antarctica).
- *Average dwelling size/average floor per capita*: these values are missing for some countries (see deliverable D5.3 Section 2.2). For all the countries/regions where these numbers are missing there is no way to compute the economic exposure, because even if the dwelling/building counts are available by using the dwelling/building fractions and the population counts, the size of the buildings or dwellings is not.
- *Replacement costs*: The replacement costs that have been introduced in most of the countries are "default" ones, computed using the procedure proposed by Geoscience Australia (see http://www.nexus.globalquakemodel.org/ged4gem/posts/ged4gem-building-replacement-cost-methodology) and based on the country GDP. In this case, the cost values are the same for all building typologies, and this is a further simplification.

6 How users will be able to provide data to GED

This section includes a short overview of the methods to ingest data into GED. These methods do not include all of the techniques that GED implemented, but only those that will be available to the general users who want to provide data to the GED.

6.1 Collecting data using the Inventory Data Capture Tools (IDCT) software suite

The Inventory Data Capture Tools (IDCT) focus on the creation of building exposure data to be easily loaded into the GED, as a Level 2 or Level 3 dataset (see http://www.globalquakemodel.org/resources/use-and-share/tools-apps/). The IDCT tools are designed to develop GEM data through any combination of the following data sources: data collected in the field, building footprints (with or without height) extracted from remotely sensed imagery, and expert opinion informed by the literature, regional reports, or an analysis of building data. The IDCT are not designed as an ETL (Extract, Transform, Load) platform for mapping and importing individual buildings or census data. Building data is collected worldwide in innumerable formats, and process of "mapping" and transforming these data into the GEM building taxonomy for risk assessment takes considerable expertise and is unlikely to be facilitated by automation. Next section provides a basic introduction to the IDCT tools at it pertains to the creation of GEM consistent data. To use the tools, the reader is advised to seek IDCT user documentation (available from http://www.globalquakemodel.org/resources/publications/technical-reports/). The following section discusses the importance of mapping schemes, as well as the process of applying mapping schemes to the database.

6.1.1 Overview of Collecting Data with the IDCT Software Suite

There are three main components of the IDCT suite of tools, as illustrated in Figure 6.1. Remote sensing tools and protocols are used to process imagery to derive building footprints. With field tools, teams characterize building attributes using a series of data collection forms. The third component, the Spatial Inventory Data Development (SIDD) tool combines these data sources and uses a statistical approach to develop a GEM dataset. Depending on the data available to the user, the tool creates building data from a combination of statistical structural characteristics and aggregates building data (number and square footage) derived from remotely sensed data or provided through the GED.

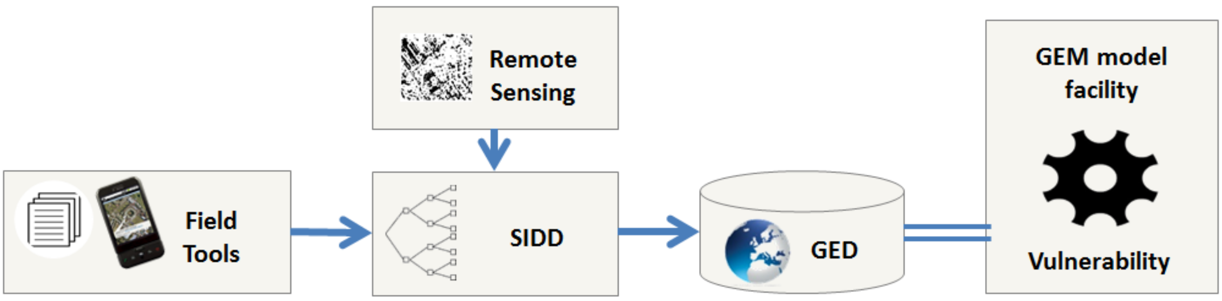


Figure 6.1 IDCT Workflow for Importing Data into the GED.

6.1.2 *Updating Level 1 data*

Updating GED data at the national level through the GED can be accomplished in the IDCT tools by updating the default mapping schemes and recreating a GEM compliant exposure database. In most cases at the national or wide regional level, there will not be resources to collect or process remote sensing data or collect data with field tools. This process is illustrated in Figure 6.2. The user begins with default Level 1 data from the GED and default mapping schemes from the OpenQuake platform using the SIDD tool. The users will then update the mapping schemes, apply them to the building data to develop an alternative distribution of structure types, and submit the data to GEM for consideration for the GED.

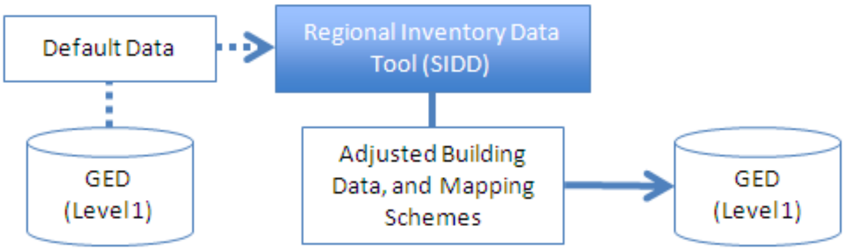


Figure 6.2 Developing Level 1 data with IDCT tools

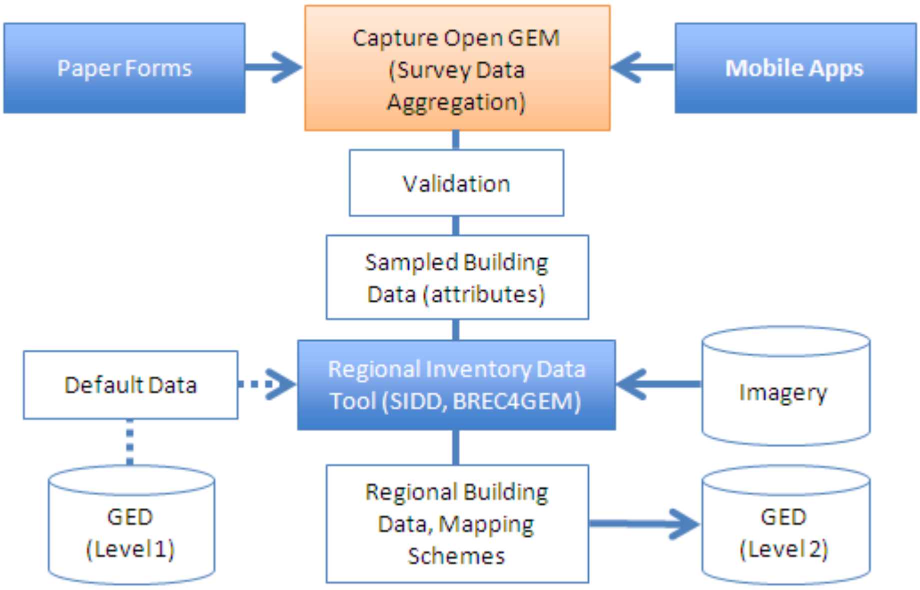


Figure 6.3 Creating Level 2 data with the IDCT tools.

6.1.3 *Populating Level 2 data using IDCT Tools*

When using IDCT to develop Level 2 data, users will have the option of creating a GEM compliant exposure dataset through a variety of techniques depending on resources available. Users may opt to modify mapping scheme data as discussed above, or they may opt to create a more robust statistical assessment of building type by implementing the GEM sampling strategy and collecting GEM taxonomy attributes with the android or windows-based field collection tools or paper forms, as illustrated in Figure 6.3. When field data is collected, a rigorous sampling strategy based on IDCT protocols must be developed and strictly adhered to.

Once data is collected, it is verified and ported to SIDD. Users may opt to use GED default square footage data and building count data, provide alternate aggregate data, or provide estimates derived from remote sensing data. If field data is collected, it can be used to develop mapping schemes within SIDD. Mapping schemes are applied to the building stock within SIDD to develop Level 2 data that is exported either in the original geographic unit provided (land use area, census tract), or by GED-consistent 30 arc second grid. Once applied, the results need to be carefully verified by the user.

6.1.4 Populating Level 3 data using IDCT Tools

Level 3 data is building level data. In circumstances where a limited geography is being considered, data is being collected post-event, or for targeted vulnerable or high-value facilities, collection of data for individual facilities users may opt to collect building specific data. In these cases, the collection of data using IDCT does not require a sampling strategy. Figure 6.4 presents the workflow. Users collect field data using mobile applications or paper forms. Building specific data can then be then loaded into the GED by GEM, or in the case of post-event data, the GEMECD (global earthquake consequences database).

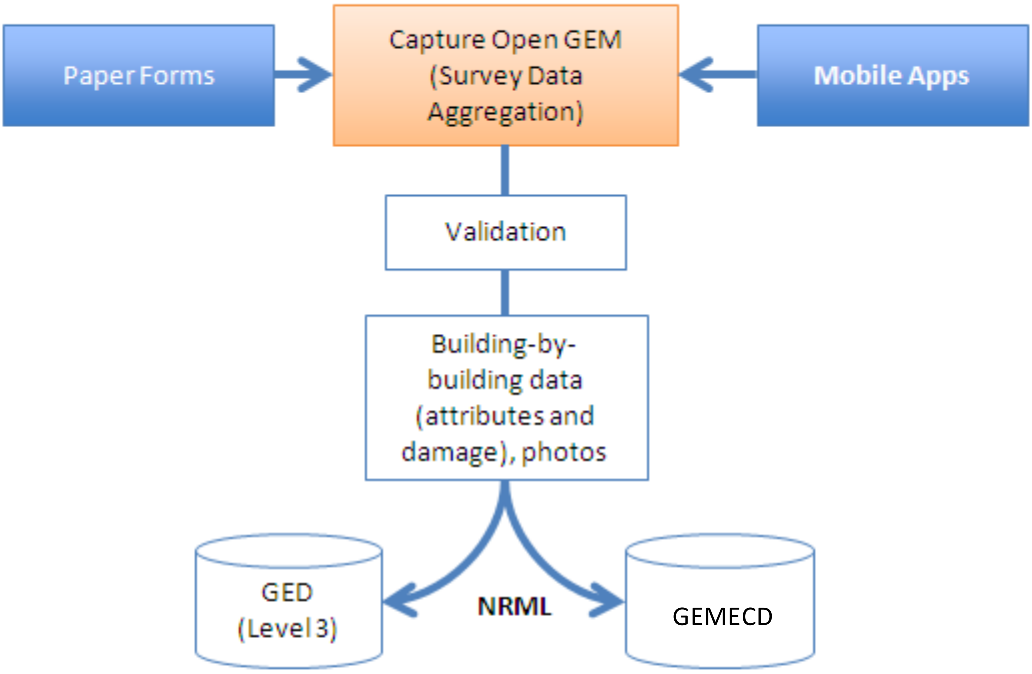


Figure 6.4 Populating the GED with Level 3 Data

6.1.5 Mapping Schemes

GEM mapping schemes track the percentage of structures associated with a given building class and a given unit of geography. For example, a mapping scheme may distribute buildings in urban areas for all of Italy into the probable building types. Such data, for certain purposes, would be considered quite crude. For example, even without the collection of field data, one may be able to assume that the construction practices in the North are very different from those in the South, and that these differences will impact vulnerability. Furthermore, those with knowledge of building practices may be able to “hazard a guess” as to what these differences are in terms of percentage of the building stock associated with a given building type. With this information, an educated user has the option to create a new study region for Italy- with separate mapping schemes for the North and South of the country. This can be achieved either through the SIDD (Spatial Inventory and Damage Data management tool), or through manipulating custom data and inserting it into the database. This section provides a general draft overview of both procedures. Final implementation will follow the same steps, but be facilitated by scripts and tools. Details of using the SIDD tool are provided under separate documentation and are provided here for context.

6.1.6 Creating a mapping scheme with SIDD

To develop or modify an existing mapping scheme in SIDD, apply it to a region, and check it into the software a user must first obtain the SIDD software (http://www.globalquakemodel.org/resources/use-and-share/tools-apps/). With the SIDD software, users have access to default mapping schemes for every country which are based on USGS PAGER. Given the use case described, where a given user wants to adjust data for a wide region, very little data is required. However, it should be noted that SIDD provides the functionality to allow users to create mapping schemes from field surveys as well. To apply the changes to Italy, users will start the SIDD program and "check out" the area of interest (Italy). They will then need to supply a GIS file with two shapes- one for Northern and one for Southern Italy defined as they see fit. The user is then presented with the mapping scheme tree. The mapping scheme tree breaks down the structural class into multiple branches, such as material, material technology, lateral force resisting system, height, and so forth. The user is then able to assess the percentage breakdown of each of these breakouts independently. One can envision this as the branches in a tree, where the leaves represent the final class (where the leaves represent all valid combinations). Once done, these statistical assessments of the branches are applied to each region of Italy, and a GED-compatible exposure database is made. Users can then upload the mapping scheme to OpenQuake, and use the resulting exposure database either by loading into a local version of the GED, or working with GEM to augment the GED. In this case, the data will mirror the format requirements of the GED.

SIDD is designed for merging field data, remote sensing data, and/or user expertise into a GED compatible database. However, there may be many users who want to get data into the GED that do not fall into this category. For example, perhaps a user is starting with their own assessment of exposure in USGS PAGER, HAZUS, RMS RiskLink, or AIR UNICEDE format. Or perhaps a user is starting with detailed tax assessor data or housing census information. In these circumstances it may be possible to glean information from the data for SIDD, but the user may wish to map the existing taxonomies to the GEM building taxonomy and apply the "leaf" statistics directly. With some "taxonomies" represented in tax assessor or census data, quite a bit of engineering expertise may be required, and breaking down the end member structure classifications may be a impediment, rather than a helpful feature. In these cases, the user will want to ingest the mapping scheme directly into the GED- either a local copy or working with authorization from GEM.

6.1.7 Inserting a mapping scheme into the GED

In order to ingest a mapping scheme separating Northern and Southern Italy into GEM, a user will need to have access to PostgreSQL administrative tools and administrative privileges. The user will follow these general steps for both Northern and Southern Italy:

1) Start a transaction
2) Insert a new mapping scheme source record

Table 6.1 Mapping Scheme Source.

Source	Date_created	Data_source	Data_source_date	Use_notes	taxonomy
Sample	2012-03-29		2010	New Mapping scheme Northern Italy	GEM

3) Retrieve the mapping_scheme_src.id for inserted record

4) Insert mapping scheme records using the id in the field mapping_scheme_src_id

Table 6.2 Mapping Scheme

Mapping_scheme_src_id	ms_class	ms_value
999	MR+RW+STRUB+MOM/LWAL////	0.7
999	MUR+CLBRS/LWAL/RW2/FW99/H:2/YP:1939/IRN/RES99/	0.2
999	MUR/LWAL/RW99//H:2///RES99/	0.05
999	STRUB+RW+MORL	0.05

5) Commit transaction

The user would have to be familiar enough with the GEM building taxonomy to determine what class is appropriate for each element of their initial database, and they would then use these classes to determine percentages that were entered into the mapping scheme. In addition, they would have to assure that the given string is valid and associated with a vulnerability function. Because this is an open system, there is no "master list" of valid strings for validation.

6.2 Providing Level 0 and Level 1 Dwelling Fractions and National Statistics

The dwelling fractions for a specific area may be provided by any expert for his/her area of interest. The procedure includes filling a very simple form and sending it to the GEM Secretariat (http://www.globalquakemodel.org/media/storage/GEM-ContributionForm_GlobalExposureDatabase_1.docx). The data will be then scrutinized by GEM experts and, eventually, added to the GED. Moreover, the GED includes, as mentioned above, national statistics needed to compute building counts from population counts. These data may be also provided using the same form.

6.3 Providing Level 0 and/or Level 1 Data from Census/MICS/DHS Datasets

Finally, dwelling fractions from national surveys, like Census, MICS (Multiple Indicators Cluster Survey) or DHS (Demographic and Health Survey) can be obtained using the methodology developed in Section 4.1. Data may include information on the actual numbers of dwellings with any combination of wall/floor/roof materials. These numbers are provided either only at the country level or, additionally, at the sub-regional level, equivalent to GADM Admin 1/2 regions, but must be complete, i.e. must include information for all regions.

7 GED Data Visualization, Query, and Subset for General Users

This section details the tools developed by GED4GEM to make the GED data available or visible to the general users. All of these tools include a user control and management part that makes available to different user groups different functionalities.

7.1 Web User Interface and Visualization Tools

The developed interface is composed of a main window to help locate the area of interest and selection tool that allow visualizing the level 0/1 available. The interface is completely web-based: data are exposed through a web server and a Geoserver, and a web browser is required to visualize the results.

7.2 Data Query and Subset Features

Data query is currently implemented via the “Export tool” in the OpenQuake platform. It presently allows to extract, after selecting a geographical area of interest (subject to limitation in size), the GED grid with all the information available: people count, dwelling fraction and, in the future, building counts and replacement costs. The data will be provided either in a format directly useful for the OpenQuake-engine, or in CSV format for easy visualization, analysis and correction/integration by the user who might have better/more detailed/more recent data for the same area.

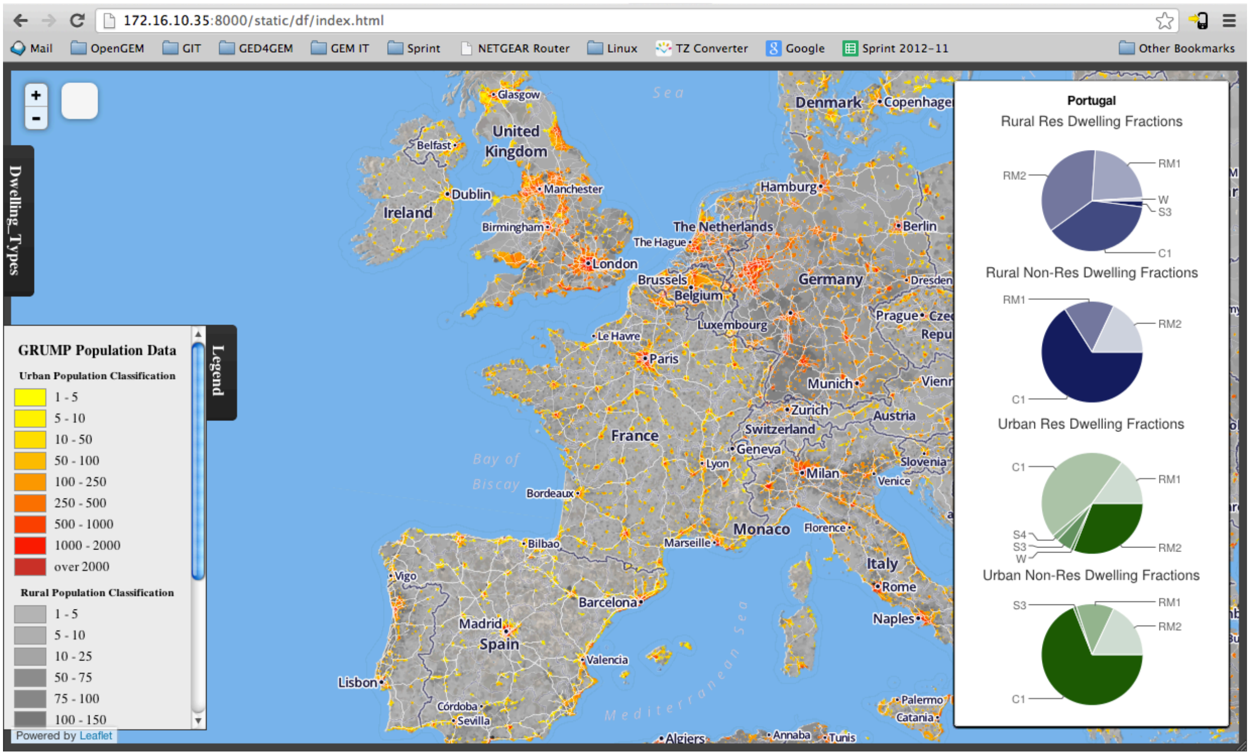


Figure 7.1 Visualization of Population Counts and Dwelling Fractions.

8 Limitations and future ideas

The most critical issues that are still open in the GED and would require some development and research work in the future are those listed here in the following.

8.1 Quality of the datasets

The need for a way to provide to the user information about the quality of the information provided by GED is a very important open issue in GEM. At the moment, each dataset in GED is time stamped and it is possible to obtain its corresponding data source, following a suggestion to the consortium by Geoscience Australia. No data is however labelled as “excellent” or “reliable” (or the opposite), because this would imply a check that must be performed by experts according to a specified protocol. Neither the protocol nor the way to evaluate this “quality label” has been decided within the time framework of GED4GEM, although this point was discussed in many meetings. There is a very demanding yet precise proposal by USGS that was not endorsed by the whole GED4GEM consortium but it is a part of the documentation of the October 2012 meeting in Pavia.

8.2 Uncertainties of the datasets

A related issue is about the uncertainties in the data. Currently the GED does not include this information, although for a specific situation it would be possible to add it. As a matter of fact, Geoscience Australia provided the standard deviation for the default replacement cost rate. However, this alone is not sufficient to estimate the uncertainty in the values in level 0/1 exposure, and thus this is left as an open issue for future research.

8.3 Time reconciliation

The datasets in GED are time stamped, as mentioned above, but there is no implemented routine to reconcile data coming from different sources in different years. For instance, at this point the dwelling fractions in one country may refer to year 2010, but the average number of people per dwelling is from 2008 and the replacement cost is computed from 2011 data. A methodology to adjust numbers to the same years by considering ancillary information which is not available inside GED (e.g., temporal sequences of GDP, or population counts) is required to obtain statistically meaningful data in areas of the world and in time periods where these numbers may be subject to large changes. Strictly speaking, however, time reconciliation can be performed outside the GED: as long as time stamped data are provided to the users, it is up to them to use them as they are or reconcile the numbers in time according to his/her best knowledge and capability.

8.4 Currency reconciliation

A similar point, but with a minor importance in the context of the GED, is the need to reconcile cost data which is available sometimes in the local currency. In GED there is no mechanism to extract the same currency in case two areas in neighbouring countries are selected and the user wants to extract exposure information for both of them at the same time. Once again, this may be considered as a problem that can be solved outside the GED, and left to the user, but there are situations (like the example above) where it may be better to have it solved by the export tools.

8.5 Rural/urban data for HAZUS and NERA

One peculiar problem in GED is due to the studies referring to whole countries or their administrative parts and providing information about buildings and/or dwelling without any distinction between urban and rural numbers. Since in GED data are stored for four different cases (urban residential, rural residential, urban non residential, rural non residential), the data are copied as they are for both the rural residential and the urban residential cases. It would be however a better option to find a way to distribute absolute numbers (such as the total number of buildings, or the total number of dwellings), perhaps according to the population ratio between urban and rural areas. This is not needed for relative numbers, such as dwelling fractions.

8.6 Non-residential areas

The survey data that are available to UN-Habitat and processed (see Deliverables D4.2 and D6.2) to be ingested in the GED included non-residential data also, but their statistical significance is dubious and therefore they were not included into the data extracted and processed and eventually ingested into the GED. It may be possible however to analyse and process these data to include information about non residential areas and make the study for these country more significant.

8.7 No level 3 ingestion script

The only way that is currently forecasted to include level 3 data into the GED is to use the direct observation tool designed by the IDCT consortium. Direct ground observations were carried out in Pavia in 2014 and they now need to be exported from the IDCT direct observation tool and imported into GED with custom ingestion scripts.

8.8 Easier ways to interact with the potential data providers

The sustainability of the GED is largely dependent on its capability to be useful, and in turn on the possibility to be easily updated and integrated with new information. Although the GED4GEM project has developed tools to ingest data available in many different formats for the different GED levels, the weak point is the interaction with people that may have (and be willing to provide) some data. So far, the interaction has been based on e-mails and personal communications. Notwithstanding the big help from the GEM Foundation, the GED4GEM consortium did not succeed unfortunately to engage the large community of potential data providers. In addition to a better “marketing” strategy, from the purely technical point of view it may be that the design and deployment of user-friendly and self-guided “input tools”, in addition to the already deployed “output tools”, is a way to improve the situation.

8.9 Data from Regional Programmes

Although there have been attempts by the GED4GEM consortium to interact with GEM Regional Programmes and receive from them some data to be ingested into GED, eventually the result is not satisfying. The data that have been received so far is very partial and most of it has still to be delivered. The GED4GEM consortium acknowledges the issues and problems in dealing with the GED schema that have been changing during the project and completely defined only a few months ago. It is true, however, that the sustainability of the GED depends on a better connection between whoever will manage the database and the users. In this sense the Regional Programmes must play a very important role, and it is expected that they will be, differently from the current situation, the primary source of data and updates to the data stored in the current version of the GED.

REFERENCES

Brzev S., C. Scawthorn, A.W. Charleson, L. Allen, M. Greene, K. Jaiswal, and V. Silva (2013), GEM Building Taxonomy Version 2.0, GEM Technical Report 2013-02V1.0.0, 188pp., GEM Foundation, Pavia, Italy, doi: 10.13117/GEM.EXP-MOD.TR2013.02

Jaiswal K.S., D.J. Wald, and K. Porter, K. (2010), A global building inventory for earthquake loss estimation and risk management, Earthquake Spectra, Volume 26, No. 3, pages 731–748

THE GLOBAL EARTHQUAKE MODEL

The mission of the Global Earthquake Model (GEM) collaborative effort is to increase earthquake resilience worldwide.

To deliver on its mission and increase public understanding and awareness of seismic risk, the GEM Foundation, a non-profit public-private partnership, drives the GEM effort by involving and engaging with a very diverse community to:

- Share data, models, and knowledge through the OpenQuake platform
- Apply GEM tools and software to inform decision-making for risk mitigation and management
- Expand the science and understanding of earthquakes.

The GEM Foundation wishes to acknowledge the following institutions/ organizations, for their contributions to the development of this report:

- Center for International Earth Science Information Network (CIESIN), USA
- Global Urban Observatory (GUO) of UN-HABITAT, Kenya
- ImageCat Inc., USA
- Joint Research Centre of the European Union (JRC), Italy
- US Geological Survey (USGS), USA
- EUCENTRE, Italy
- Geoscience Australia (GA), Australia

GEM Foundation
Via Ferrata 1
27100 Pavia, Italy
Phone: +39 0382 5169865
Fax: +39 0382 529131
info@globalquakemodel.org
www.globalquakemodel.org

DECEMBER 2014

GEM
GLOBAL EARTHQUAKE MODEL
working together to assess risk