SIBYL

(SeIsmic monitoring and vulnera**B**ilit**Y** framework for civi**L** protection)

Agreement number: ECHO/SUB/2014/695550

Deliverable DB3: Guidelines of the mobile mapping system and remote rapid visual screening

Version 1.0 February 2016

Project start date: 01.01.2015 End date: 31.12.2016

Coordinator: Prof. Dr. Stefano Parolai
Centre for Early Warning Systems
Helmholtz Centre Potsdam GFZ German Research Centre for Geosciences, Potsdam, Germany

Contents

Introduction

A key concept in Disaster Risk Reduction (DRR) is the importance of relevant, reliable, and up-to-date information. Being able to collect exposure information efficiently, at multiple geographical and temporal scales, allows risk practitioners to move towards more sustainable assessment schemes. The relevancy and reliability of the information to be collected depends on the particular approaches selected for data sampling, collection and analysis. The timeliness of the collected information depends on the resources available, but also on the efficiency of the overall operational procedure.

Several application scenarios can be considered, for instance, where the above-mentioned constraints play a critical role:

Scenario A – induced seismicity

The abrupt appearance of small, but perceivable seismicity, potentially connected with anthropogenic activities (e.g., fracking operations) concerns the population of a small town, located in an area not known to be exposed to seismic hazard. Civil protection authorities are called upon for prompt action, but no information on the expected seismic vulnerability is available.

Scenario B – uncontrolled urbanization

A large town in an economically developing country is subject to rapid and uncontrolled urbanization (urban sprawling) which has radically changed its exposure and vulnerability to different natural hazards over the last 10 years. Risk practitioners and civil protection authorities are subjected to increasing pressure to undertake cost-effective prevention and mitigation actions.

Scenario C – post-disaster assessment

A significant earthquake has just affected a heavily-urbanized region, generating substantial damage. Civil protection authorities and first responders need a timely and reliable preliminary assessment of the amount and spatial extent of damage, in order to better plan the emergency management and the first recovery actions.

In all of the mentioned scenarios, rapid and efficient action is needed, with the aim of collecting exposure data within an urban environment and analysing its vulnerability with respect to a possibly imminent threat. The REM (Rapid Environmental Mapping) platform has been developed by GFZ and further advanced within SIBYL, in order to provide a comprehensive set of tools and methodologies that aim at streamlining the exposure modelling process for different types of application scenarios.

This document describes in detail how to employ the REM platform to rapidly collect and integrate vulnerability attributes of the built environment. The REM strategy and the corresponding tools are presented, and discussed. In the Appendix, an introduction to the use of sampling techniques for exposure characterization is provided.

The REM data collection strategy

Modelling the exposure and vulnerability of a modern complex urban environment is a challenging task which entails collecting information at different spatial scales. Some of the features we are interested in, such as, for instance, a building´s roof type, might be observed by looking at a very high resolution satellite image (or high-resolution orthophotograph). Other features, such as the number of storeys of the building or its use require a different perspective, and a 'street-view' picture would be more suitable. In other cases, a more in-depth analysis might be conducted (for instance, in order to precisely evaluate the load-resisting system of the building), therefore requiring the closer physical proximity of a skilled expert.

The SIBYL project has further advanced a data collection strategy, called REM (Rapid Environmental Mapping), exploiting mobile mapping systems combined with remote sensing and geostatistical analysis. The REM is a set of methodologies and tools for carrying out efficient and reliable surveys aiming at characterizing the physical exposure of an urban environment. The underlying data collection strategy builds upon several key concepts:

- Large-scale data extracted from remote sensing are used to have a preliminary characterization of the target environment.
- Finer-scale information has to be collected in-situ, with a statistical sampling approach.
- Street-view observations over finer scales provide an initial characterization of the built-up environment which can be integrated with small scale information from different sources.
- The use of a mobile mapping system along optimized routes allows for a rapid collection of georeferenced visual information which can be analysed off-line and remotely.
- Further direct observation of target structures is carried out based on the results of the first characterization phase, and has to be seamlessly integrated into the overall information lifecycle.

The basic steps for carrying out a REM survey are as follows:

1. Stratification of target areas using remote sensing analysis.
2. Sampling and optimized routing.
3. Visual data collection through the MOMA (MObile MApping) system.
4. Analysis of data collected through the REM RRVS (Remote Rapid Visual Screening) web platform.

In the following sections, the four components will be individually discussed.

Stratification of target areas using remote sensing analysis

In the previous sections, the basic concepts about sampling have been presented. This section demonstrates how the same concepts can be exemplified and how remote sensing information can be exploited to achieve greater efficiency in collecting exposure-related data. The information collected *in-situ* on the ground is used for several purposes, the most important being:

- The collection of dense *in-sit*u observations of the physical, social, and economic landscape (e.g., residential buildings and their structural features),
- The collection of ground-control-points and environmental observations to constrain small-scale models obtained by statistical learning.
- The collection of ground-truth information for the testing, validation and accuracy assessment of EO-based products.

The collection of in-situ data is often challenging, requiring significant resources in terms of personnel, time and budget. The choice of a suitable sampling design will therefore help in achieving greater efficiency, allowing more high-quality data to be collected and integrated, and therefore improving the subsequent risk-related estimates.

Unfortunately often there is scarcity of ancillary information which would allow the development of sampling strategies more efficient than simple random sampling (such as, for instance, stratified sampling). In this context, remote sensing data can be exploited. In particular, Landsat imagery can be used to generate in a short time a first stratification of the considered area by subdividing it into smaller geographical entities which we suppose have a certain degree of homogeneity. A more detailed explanation of the procedure is provided in the SIBYL deliverables DB1 and DB2. Here we consider the result of the remote sensing analysis for the town of Cologne, in Germany, as shown in Figure 1.

Sampling and optimized routing

The stratification obtained in the preceding section can now be used to generate an adequate sampling distribution on the ground. A set of points is generated in order to sample the geo-cells of the considered area according to their classification and proportional to their surface coverage. The resulting set of points is shown in Figure 2. The obtained sampling points can be used as input to the routing engine, together with a topologically corrected street network of the area of interest.

In order to optimize the routing, a number of sample points are randomly selected from the sampling set defined above, and used to select a related set of road segments, that will make up the path of the mobile mapping system. However, it is also important to define in which order to visit the selected road segments in order to make the data collection time- and cost-efficient. Moreover, in-situ data capturing, especially in urban areas, may be influenced by driving restrictions (accessibility, turn-restrictions, one-way streets, etc.) and cost factors (length, time, money, etc.).

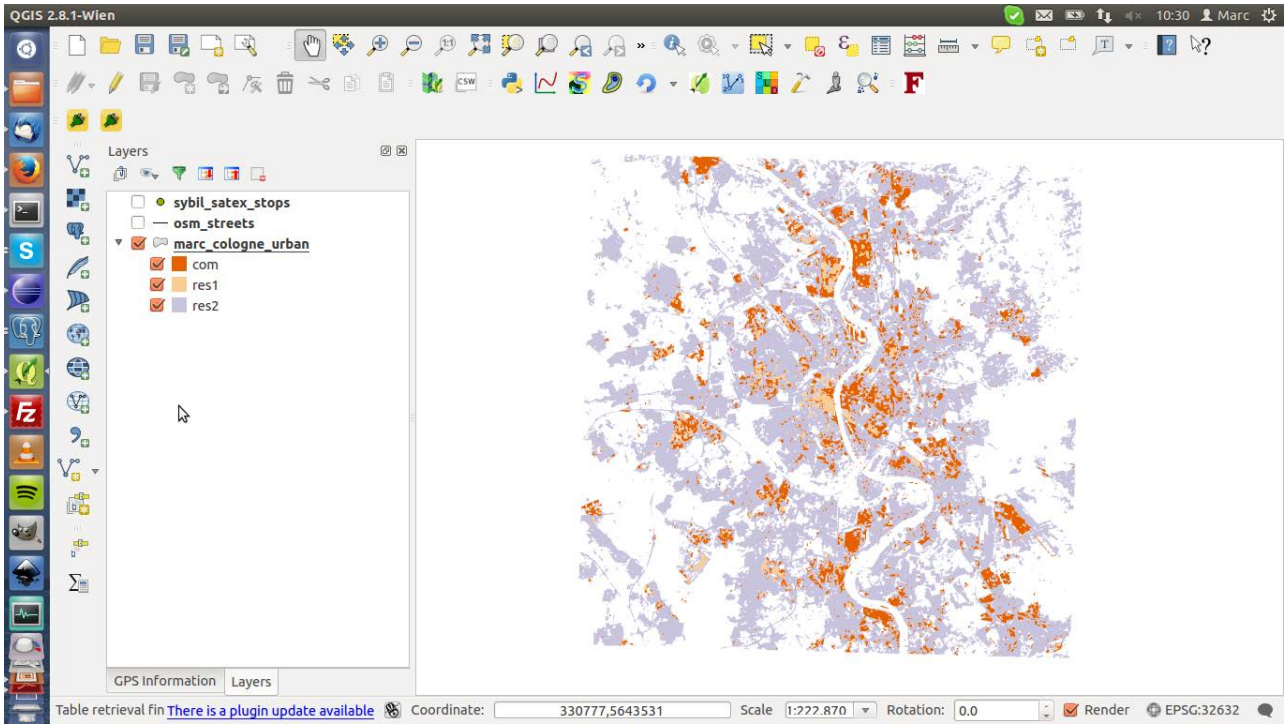

Figure 1 Results of the preprocessing and classification of a Landsat image of Cologne, Germany. Only the three classes related to urban areas have been considered.

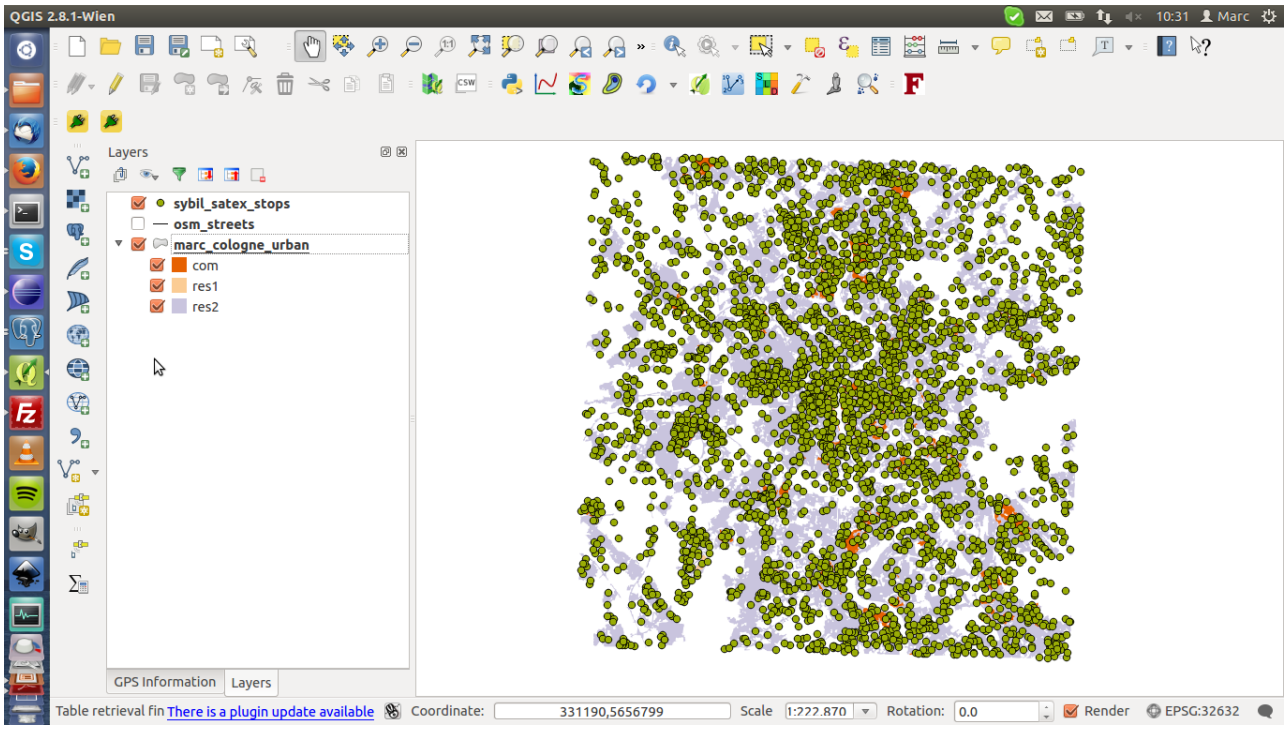

Figure 2 Generation of stratified sampling distribution, with proportional allocation, according to the considered classes.

The main steps that are involved in tackling the aforementioned challenges include the following:

- Order the sample points based on a predefined cost function, especially the start and end points of the planned route.
- Find the route through all the ordered stops that minimizes the cost function while considering the restrictions imposed by the road network.

As input data for the routing operation, a road network dataset has to be provided. Such information may be available from qualified institutional sources, but often as a simpler alternative the data provided by OpenStreetMap (OSM[1]), as the one shown in Figure 3, can also be used.

The data needs to be topologically corrected and are used to create a routable geometric network with defined cost-factors for travelling along street segments. The cost-factor used within a standard routing operation is the length of a street segment. Additional cost-factors and restrictions, such as street quality, turn restrictions or traffic information, can be added to the network if available for an area of interest.

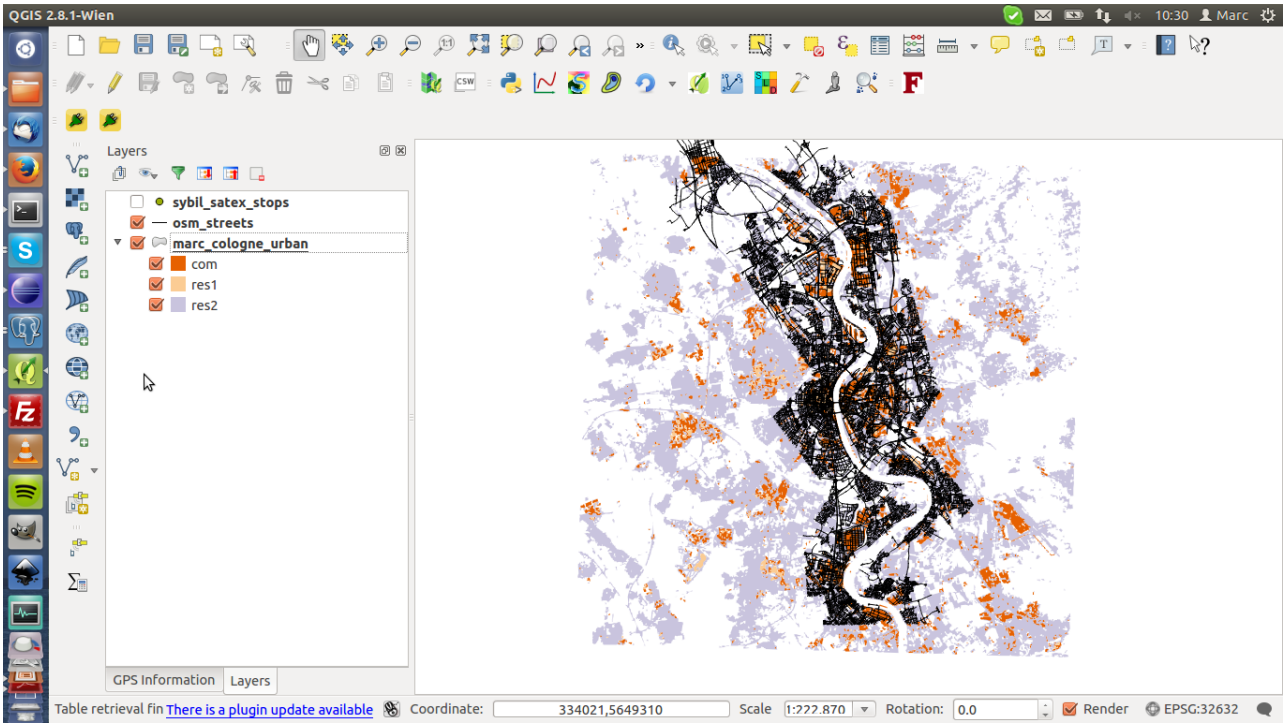

Figure 3 Road network obtained from OpenStreetMap (OSM) for the city of Cologne, Germany.

The actual routing problem can be reduced to the common Travelling Salesman Problem (TSP). The TSP is a well-studied combinatorial optimization problem in which a traveller is required to minimize the total traveling costs in order to visit all the stops on his list only once. To solve the TSP, a routing engine can be implemented directly with the database (server-side). The routing engine is based on the pgrouting[2] extension to PostgreSQL[3] and implements a set of custom functions for advanced routing operations. The functions include, amongst others, solutions for the

[1] www.openstreetmap.org/
[2] http://pgrouting.org
[3] http://www.postgresql.org/

TSP under the consideration of custom cost functions and a multiple Dijkstra algorithm to determine the best route through a series of stops while minimizing the cost function.

In a first step, the sampling points are filtered and mapped on to the street network to define the route stops that should be covered during the field operation. The closest nodes of the street network are selected for each sampling point as route stops using a straight line distance from point-to-point. Only one route stop is selected in the case where multiple sampling points refer to the same network node. This effectively filters the sampling points based on their accessibility. Once identified, the route stops are fed into the routing engine and the TSP solver is applied, where the cost-factor to be used is defined as an attribute in the street network data. A Dijkstra algorithm is than applied multiple times between the sorted stops in order to calculate the shortest path across all the stops. Figure 4 shows and example of the route stops for the case of Cologne.

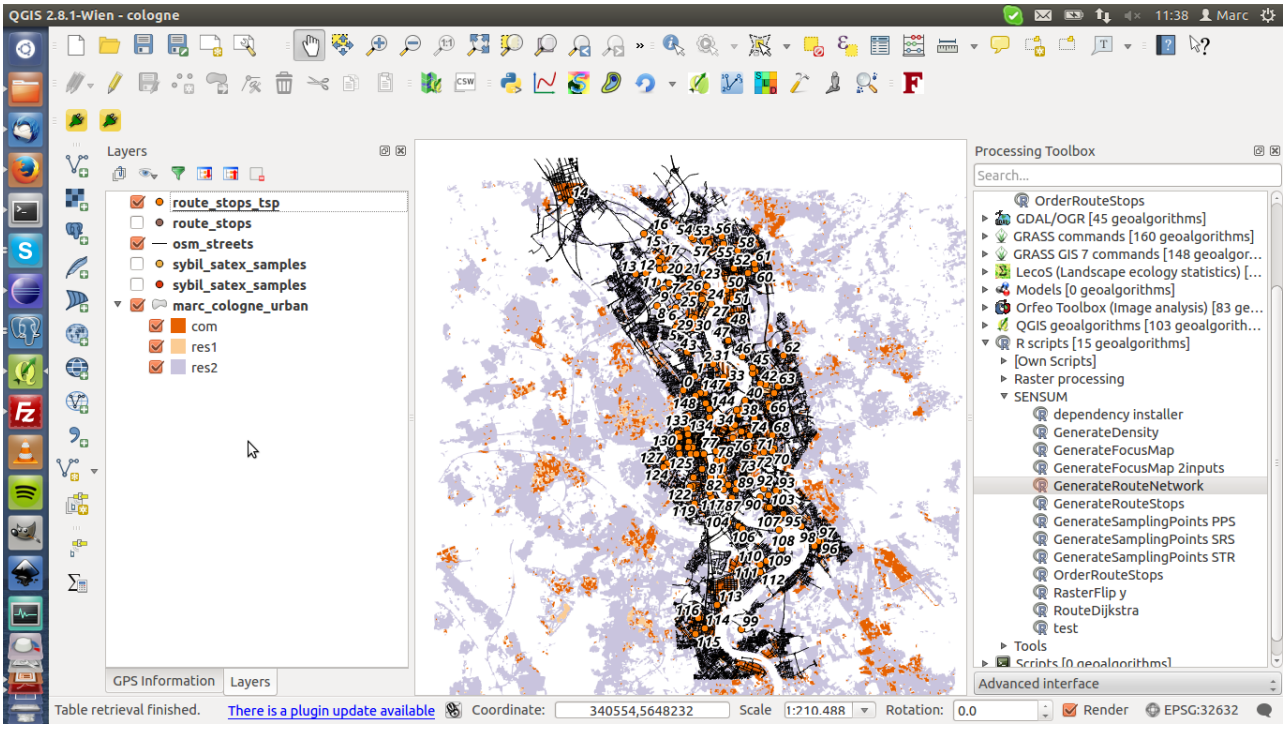

Figure 4 Route stops computed from the sampling set and the available road network for the case of Cologne, Germany.

The routing engine can be successfully used to optimize the implementation of the planned survey (that is, the coverage of the sampling units selected according to the chosen sampling design) accounting for different time and cost constraints which can significantly impact upon the survey resources. For instance, placing a penalty on the repeated scan of the same street would force the routing engine to enlarge the geographical scope of the survey, adding potentially additional useful observations to the planned ones. Also, highly dynamic parameters, such as, for instance, real-time traffic information, might be considered in the routing phase which could also be conducted *in situ* using a mobile platform. This would allow the mobile mapping system to adapt to changed environmental conditions without losing the general focus of the survey.

The final routing is shown in Figure 5. The routing engine has been implemented in a free, open-source environment by exploiting the computing capabilities of the postgreSQL/postGIS database solution.

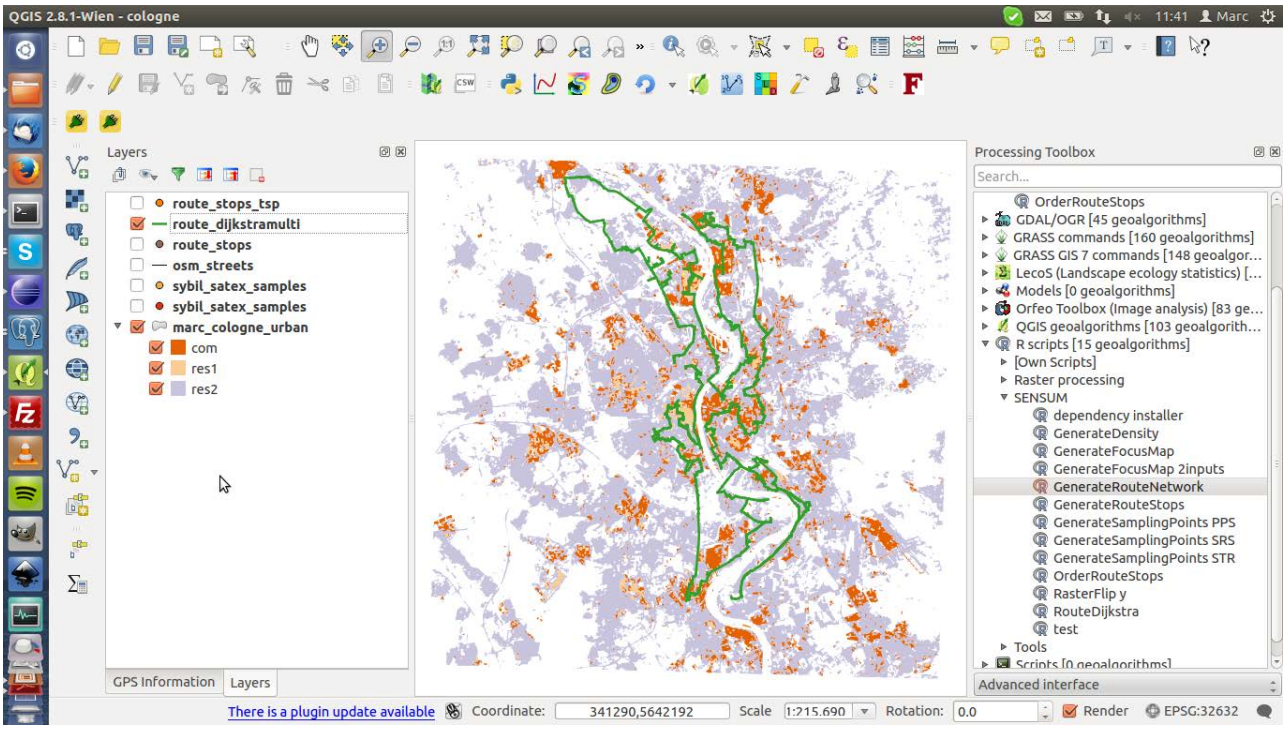

Figure 5 Final optimized route implementing the desired sampling on the ground for the case of Cologne, Germany.

Visual data collection through mobile mapping system (GFZ MOMA)

Once the scope of the survey has been clarified, a preliminary remote sensing analysis has been carried out, and the resulting stratification has been used to plan the most convenient route for observing the desired features on the ground, the actual survey can be carried out.

The proposed methodology is based on the use of a mobile mapping system, developed by GFZ (MOMA). The system is composed of an omnidirectional camera mounted on the roof of a car (see Figure 6), and integrating a GPS receiver and an optional inertial sensor. The system can be easily driven along in the targeted urban environment, capturing a dense (from 5 to 10 fps) stream of high-resolution (up to 5400x3700 pixel) geo-referenced panoramic images. An example of a captured panoramic image is provided in Figure 7.

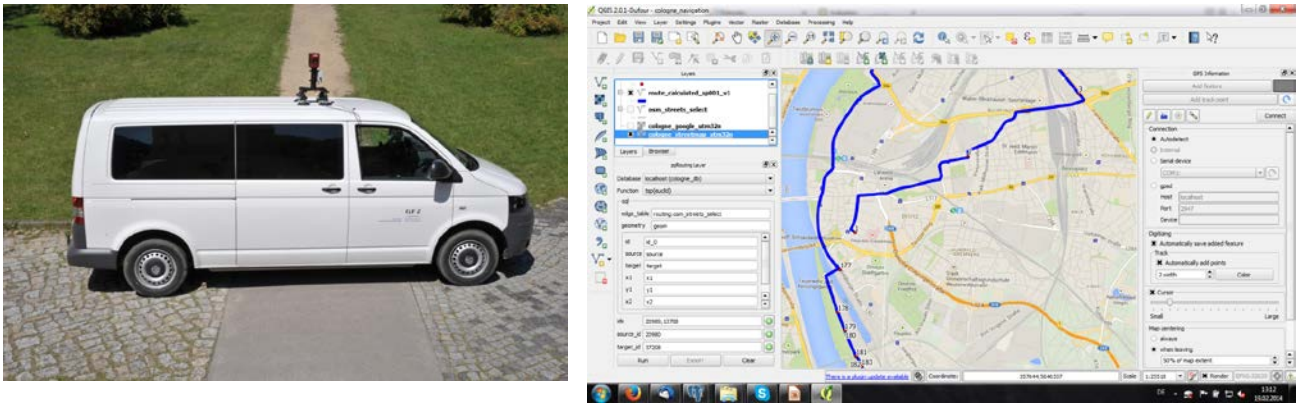

Figure 6 The GFZ Mobile Mapping system (GFZ-MOMA). Left. The installation of the MOMA system on a vehicle. Such an effort requires only around 15 minutes and does not need skilled operators. Right. The real-time user navigation interface used to operate the MOMA system.

Figure 7 Example of an unprojected omnidirectional image. The horizontal field of view is 360°, the vertical field of view is around 170°.

The vehicle is usually operated by two persons: a driver and a navigator. The navigator has the duty of assisting the driver in following the planned optimized route. In case the optimized route cannot be followed (for instance, because of unexpected traffic jams, works in progress, accidents, etc.) the navigator can choose a different route using his or her own judgement. The total adherence to the planned route is not fundamental. What is important is to comply as much as possible with the sampling distribution that originated from the particular route.

A simple real-time navigation interface can be easily realized with a laptop running the QGIS platform, and a connected portable GPS (Figure 6). As this is a full-featured GIS platform, it is possible to show on the underlying map not only the actual path of the MOMA and the planned

one, but also any other ancillary geo-information which can help the navigator to carry out their duty.

The MOMA system is battery-operated and can be driven for up to 6 hours before requiring re-charging (with an optional second battery, naturally the operation time can be increased). In a normal day (6 hours) of operation in a dense urban environment, the system can drive more than 100 km and acquire tens of thousands of geo-referenced images.

In the post-acquisition phase, the captured images are stitched and saved as equirectangular panoramic images, and a set of metadata is embedded in order to keep a record of the acquisition parameters, i.e., the geographical coordinates, the processing phases and any other information which can be relevant in the subsequent phases.

Once the survey is terminated, the images can be transferred to the REM database, described in the deliverable DB2, and analysed through the RRVS interface described in the following.

Analysis of data collected through the REM RRVS web platform

The Rapid Remote Visual Screening is a modern version of the well-known Rapid Visual Screening methodology (see ATC-13 and FEMA-154[4] methodologies), largely used in the engineering community.

The geographic locations where the images have been captured are stored in a database. A complete solution for visualization, analysis and entry of the observed data is depicted in Figure 8. A desktop operator can efficiently conduct virtual observations of the environment of interest, and store the captured attributes of the population in an efficient relational database, for later estimation and processing. The RRVS tool is described in more detail in the SIBYL deliverable DB2.

The main task of the RRVS tool is to quickly associate to building, described by its geographical coordinates or by its footprint in a GIS model, to the set of structural and non-structural features included in the particular taxonomy considered. This information can then be used in the analysis phase, to estimate the structural typology of the building, and its expected vulnerability with respect to earthquake or other natural hazards.

[4] http://www.fema.gov/media-library/assets/documents/15212

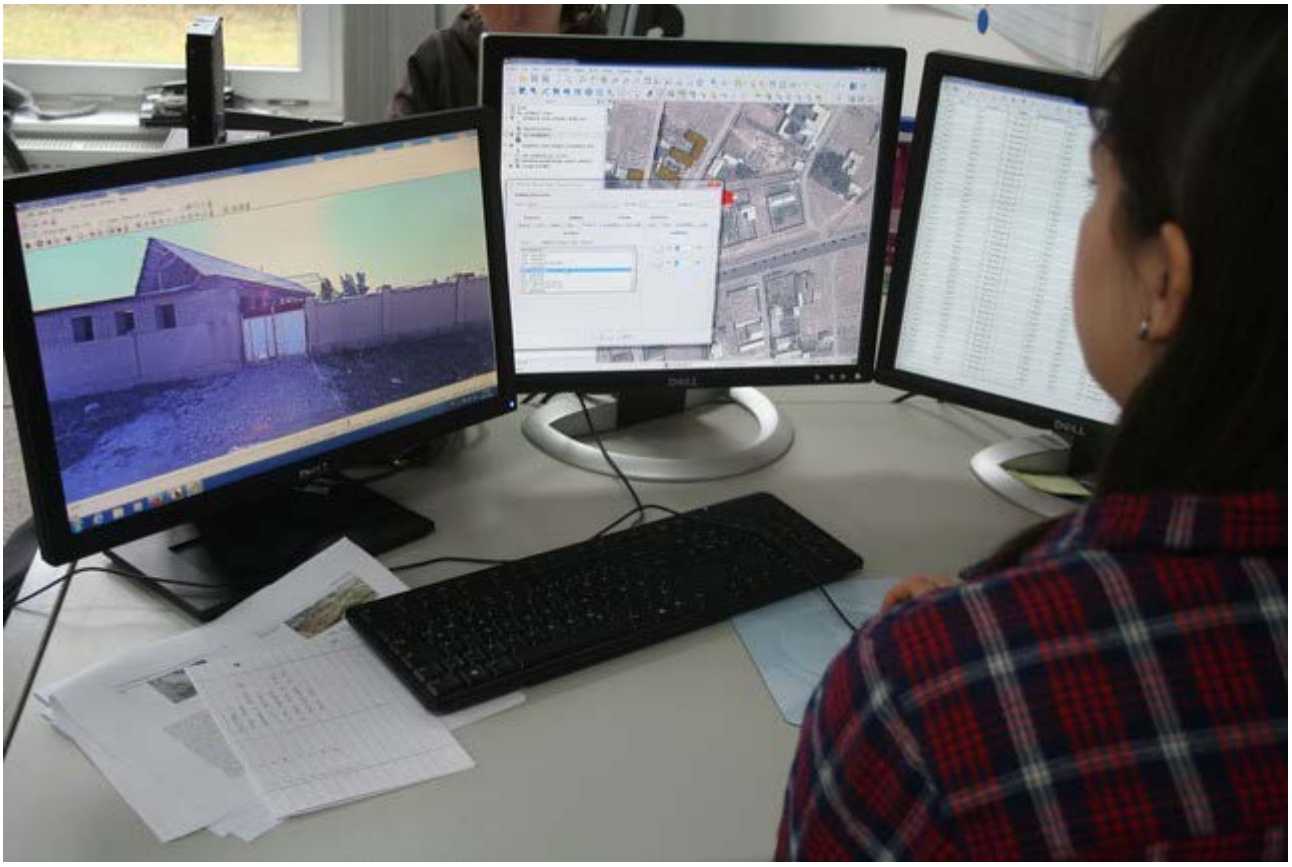

Figure 8 Desktop-based analysis of the acquired visual information (Remote Rapid Visual Survey).

In order to accomplish that, a web-based platform (see Figure 10) has been developed within the project SIBYL, for the remote, rapid screening of the buildings. The system can be accessed by remote through its public access point. The users have to provide a user name (previously registered into the system) and a task number. The task number picks up a subset of the buildings to be inspected, which are previously selected from the database according to the specific sampling schema to be realized. Every task is composed of a variable number of buildings to be inspected (e.g., 100 units). The spatial distribution of the buildings composing the task can vary according to the sampling approach. The use of tasks allows several operators to work in parallel on the same dataset, therefore increasing the flexibility of the system and its potential applications.

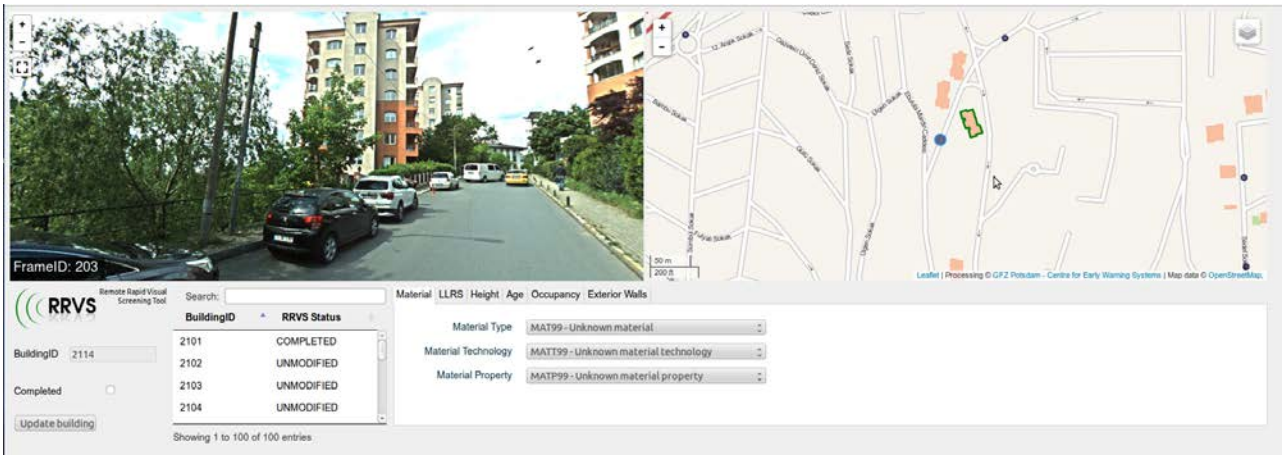

Figure 9 Web-based interface of the RRVS (Remote Rapid Visual Screening) Tool.

This tool, combined with those presented in the previous sections, allows for the implementation of rapid, inexpensive surveys of the urban environment which have the two-fold advantage of:

1. Providing reliable first-order estimates of the attributes of interest and thus allowing the development of more realistic vulnerability and risk models, as well as providing a basis for further data collection activities.
2. Collecting in a short time a valuable visual database of omnidirectional images, georeferenced and relevant for particular areas of the considered urban environment). These images might thus be used in subsequent RRVS activities to collect other attributes of interest, and be related to other reference hazards (such as winter storms, for instance).

Conclusions

In this deliverable, the REM (Rapid Environmental Platform) and the related data collection strategy has been introduced. Such tools and activities aim at exploring and assessing the potential integration of planned routes, GIS systems and mobile mapping technologies in pre- and post-disaster monitoring and assessment tasks. Particularly in extended urban environments, the use of REM would strongly improve the operational capabilities of most of the interested end-users.

The proposed methods are comprised of two main components:

1) A design and optimization platform for the data collection surveys to be carried out by the mobile mapping system. This is composed of a dedicated database, and by different processing tools.

2) The mobile mapping system (GFZ-MOMA), devoted to the actual collection of georeferenced visual data,

Several statements with regards to this framework can be made which are of general validity:

- A simple, customized mobile mapping system can efficiently and rapidly collect large amounts of high-quality, georeferenced visual data about building stock, infrastructure (e.g., roads) and in general the features contained in the visual observation of the surroundings which relate to the management of insurance policy claims in case of an earthquake. Such a system can be easily deployed wherever and whenever it is needed, does not require a particularly high level of skill to be operated, and is economically viable.
- The data collected through the mobile mapping system can be easily distributed and analysed remotely, and the intensity of the analysis (aka the number of remote operators allocated to the analysis) can be easily scaled up or down according to the contingent necessities and on the business case. For instance, additional skilled operators could be seamlessly integrated into the analysis protocol in case a bigger effort is requested on a temporary basis, for example following an earthquake.
- By applying a consistent prioritization and optimization approach, a strong increase in the efficiency of the survey can be obtained, especially when small sample sizes are required in order to decrease the impact of the survey on the available resources.
- Automatic routing optimization proved to be a simple and scalable solution to optimizing the actual data collection in urban environments. The implementation of the routing engine server-side moreover allows for the implementation of compact, flexible and reliable IT frameworks for data collection, integration and access.
- In order to maximize the reliability of the system, auxiliary information should be collected and kept up-to-date in order to provide a consistent reference for prioritization and optimization of the survey activities. The collected data would moreover have a significant value, since a better description of the exposure and vulnerability of the building stock could be achieved by, for example, analysing the collected data and integrating them into a flexible geographical database.

- Pre-event and post-event surveys can be independently optimized by integrating multiple data (possibly including statistical inferences), thus resulting in a consistent improvement in the operational capabilities in the field of data collection and related analytics.
- With REM, a new generation of iterative and incremental data collection and integration approaches can be realized. This would provide end users interested in vulnerability monitoring and risk assessment with a powerful and scalable tool to increase the reliability and timeliness of the resulting model, and to undertake more informed mitigation and prevention activities.

Appendix A - Anatomy of a survey for exposure assessment

We define a survey as any activity that collects information in an organised and methodical manner about characteristics of interest from some or all units of a population, using well-defined concepts, methods and procedures, and compiles such information into an actionable structure.

Every survey begins with the need for information, where no data (or insufficient data) are available, or their reliability is deemed not acceptable by the end-users.

A typical example within the framework of seismic risk reduction is the need for consistent information about the seismic exposure, that is the ensemble of all assets that are exposed to a certain level of seismic hazard and are susceptible to be damaged, thus generating a certain amount of loss. This ensemble, or model, usually includes (but is not limited to) people, buildings and infrastructure.

Throughout this document we will take as a working example the collection of a hypothetical exposure model for a medium-sized town.

In the following sections we will briefly review the main phases a survey can be broken into, within the particular focus or our example application.

A survey must be carried out following precise procedures, if the results are to yield accurate and meaningful information.

The main steps a survey is composed of are:

1. Formulation of a statement of objectives.
2. Selection of the survey framework.
3. Determination of the sampling design.
4. Data collection.
5. Editing and imputation.
6. Estimation.
7. Data analysis.
8. Documentation.

A brief description of each step follows.

Formulation of Statement of Objectives

A very important task in a survey is to formulate the **Statement of Objectives**. This establishes the main information needs, as well as the operational definitions and the analysis plan. This step of the survey determines what should be included and what not, or otherwise stated, what the end-users need to know versus what would be useful to know.

Developing the Statement of Objectives is a multi-step procedure, involving primarily the end-users and the parties physically implementing the survey. The following points should be carefully addressed:

- The information needs.
- The end users and the use of the collected data.
- Concepts and operational definitions.
- The survey content and analysis plan.

The information needs (stating the problem)

The first step is to describe in broad terms the information needs of the end user(s).

Take as an example the case where the Ministry of Emergency Situations of Kyrgyzstan decides to carry out a seismic risk assessment for a large town (say the capital, Bishkek) in order to evaluate possible prevention/mitigation actions to be implemented in the next 5 years.

The assessment of risk includes the evaluation of the seismic hazard for the considered target (in our example, this is provided by the Kyrgyz Institute of Seismology) and the implementation of a suitable vulnerability model, whose input is the exposure model. Since the risk assessment is focusing mainly on the social consequences of a possible damaging event (casualties and fatalities among civilian population), the main focus of the survey can therefore be described as:

The collection of data for an exposure model for the city of X, limited to the set of residential buildings and the people living therein. The survey should focus on the characteristics of the buildings relevant for the assessment of their seismic vulnerability.

This clarifies the broader scope of the survey.

The users and the use of collected data

The end-users of the data being collected are the parties actually using the data for the completion of the proposed, overarching activities. This should not be confused with the client of the survey (practically the party which is covering the expenses), even if this is often the case.

In our hypothetical case, the civil protection authorities are going to perform the final assessment via their GIS technical office workers, who are thus the real end-users.

Determining the correct end-users is paramount, since they should always be involved in the planning phase and should closely follow the development of the data collection activities themselves.

It is also very important to individuate all groups who should be involved in the planning phases (even if not being strictly end-users) in order to focus on all important aspects of the survey. For instance, in our example, since the data on the exposure will serve as the input of a vulnerability assessment, a group of structural engineers should be involved in the first planning phases to better address the information needs.

Therefore, the end-users group will include:

- Civil Protection authority (CP);
- GIS technical office of the CP;
- Engineers and consultants on vulnerability assessment.

Concepts and operational definitions

In order to identify the data required to meet end-user´s objectives, the party carrying out the survey needs clear and precise definitions.

To the extent that is possible, recognized standards should be used. This facilitates the communication with the end-users and ensures consistency across surveys. Moreover, this allows for a more efficient analysis and sharing of the collected data.

In order to properly define the operational definitions, clear answers have to be provided to the following questions:

i. What?

ii. Where?

iii. When?

i. What (or who) are the end-users interested in?

Following our example, what characteristics of the buildings are the end-users interested in? Height (or number of floors), date of construction, material of the walls, material of the lateral-load-resisting-system, type of roof, etc. are often attributes of interest. Every single attribute should therefore be clearly defined, possibly using already accepted standards.

For more consistent and reusable results, a suitable taxonomy should be used. Several taxonomies have been proposed in the seismic risk assessment community, including, for instance, those from HAZUS[5], EMS-98[6] and GEM (Global Earthquake Model)[7]. The REM platform proposed in the SIBYL project makes use of a taxonomy based on the one proposed by the GEM, which can be easily extended to different natural hazards.

ii. Where are the units of interest?

This refers to the geographical location of the units to be surveyed, or better, to the geographical composition of the sampling framework. In the case of our example, the boundaries of the selected town have to be agreed upon: is the municipality providing the spatial boundary? Or is the metropolitan area? Or just the centre? Is there any area to be excluded (perhaps because of its industrial or military nature)?

[5] http://www.fema.gov/hazus
[6] http://www.franceseisme.fr/EMS98_Original_english.pdf
[7] http://www.globalquakemodel.org/

iii. What is the reference period of the survey?

What time period should the collected data refer to (*when*)? In the considered example, the survey should take into account buildings of any age, taking a snapshot of the state of the built-up environment as up-to-date as possible. In a slightly different context, a survey could be devised focusing only on buildings built before a defined period, for instance in order to verify the compliance of the structures with respect to the related building code.

In addition, other concepts could be defined which relate to the objective of the survey. For instance, a set of demographic concepts such as the following could be clarified:

- A dwelling is any set of living quarters that is structurally separate and has a private entrance outside the building or from a common hall or stairway inside the building.
- A household is any person or group of persons living in a dwelling. A household may consist of any combination of: one person living alone, one or more families, a group of people who are not related, but who share the same dwelling.

The survey content and analysis plan

The previous concepts should allow a fairly general, but clear level of clarity in the identification of the needed information. If necessary, the content of the survey has to be further specified and agreed with the end-users, often in an iterative process, until every aspect of the survey is covered.

In the considered example, for instance, we already stated that resorting to a standard taxonomy for defining the buildings´ attributes would simplify the task of describing in detail the general object of the survey. In this phase, the buildings´ attributes should be screened in order to identify the attributes that should be collected in the field, and those that could be inferred or extracted from other types of available information.

Moreover, since a socio-economical component should also be addressed in the survey in order to determine the number of people that live in the considered buildings, an additional set of attributes has to be specified and included in the survey.

Once all the items to be measured have been identified, the next task is to determine how much detail is required for each item, and the format of the results. What measures, counts, indices, etc. are needed? Are estimates for sub-populations required?

Referring to our exemplification for instance, is it necessary to distinguish between different age groups of the surveyed buildings? In such a case, what age groups should be chosen, for example, the ages corresponding to the application of different building codes? Should continuous or categorical data be used? For instance, is the height of the buildings to be considered a continuous variable, or would a categorization (low-rise, mid-rise, high-rise) suffice, and in this case, how to define the different categories?

Selection of a Survey Frame

The survey framework provides the means of identifying the units of the survey population. The framework can be either:

- A **conceptual list**: for instance, a list of all administrative buildings built before 1970, or
- A **geographic list** in which the units of the list correspond to geographical areas (also indicated as geo-cells) and the units within the geographical areas are buildings, or single households, businesses, etc..

Determination of the Sampling Design

Two kinds of surveys can be implemented: census surveys and sample surveys. In a census survey, data are collected for all units in the population, while in a sample survey, data are collected for only a fraction (typically very small) of the units of the population.

In turn, two types of sampling exist: *non-probability* sampling and *probability* sampling. Non-probability sampling provides a fast, easy and inexpensive way of selecting units from the population, but uses a subjective method of selection.

In order to make inferences about the population from a non-probability sample, the data analyst must assume that the sample is representative of the whole population. This is often a dangerous assumption given the subjective method of selection.

Probability sampling is theoretically more complex, and resource-intensive, but since units from the population are randomly selected and each unit´s probability of selection can be calculated (or at least estimated), a reliable estimate can be produced along with estimates of the sampling error, and hence inferences can be made about the overall population.

The REM methodology is based on probability sampling, since this is the most efficient approach to exposure characterization when wide geographical areas are targeted, and the number of structures to be investigated is too high to be tackled with a census-like approach.

The sampling design represents one of the most relevant parameter of the probability sampling. The sample design depends on such factors as: the survey framework, how variable the population units are and how costly it is to survey the population, etc. The sampling design partly determines the sample size, which impacts directly on survey costs, the time required to complete the survey and other important operational considerations.

Determining the sample size is usually a process of trying to fulfil as many requirements as possible, one of the most important being the quality of the estimates, as well as being an important operational constraint.

In this document we will consider how this difficult process can be eased up by exploiting a favourable combination of earth observations and smart technologies for collecting data on the ground.

Data collection

Data collection refers to the actual process of gathering the required information for each selected unit in the survey. In our example, this mainly refers to direct observations and on the use of

available administrative data, but the basic methods of data collection also include the enumeration (where a responder completes a questionnaire with the assistance of an interviewer or alone).

Data collection can be paper-based or computer-assisted. Paper-based methods are still largely used, but present several inconveniences, including the need for post-collection processing activities in order to transform the collected data into a machine-readable form. Computer-assisted methods, also thanks to the current availability of powerful and affordable portable devices, are becoming increasingly popular and convenient. The REM platform, for instance, is compatible with the IDCT (direct observation tool) developed within the GEM project.

Editing and imputation

Editing is the application of checks to identify missing, invalid or inconsistent entries that point to data records that are potentially in error. The purpose of editing is to better understand the survey processes and the survey data in order to ensure that the final survey data are complete, consistent and valid. Edits can range from a simple manual verification performed by the operator during the data collection, to sophisticated automatic verification performed in a post-collection phase. The amount of editing performed is always a trade-off between getting the highest quality records and spending a reasonable amount of resources achieving this goal. Of course, the amount of required editing varies with the particular method chosen for data collection. A careful planning of the survey and the involved technologies can contribute to minimize the resources necessary to ensure a reliable gathering of data, as well as suitably trained survey personal.

Imputation is a process used to determine and assign replacement values to resolve problems of missing, invalid or inconsistent data.

Imputation can improve the quality of final data, but the particular approach chosen for imputation has to be carefully chosen, in order to avoid introducing a bias in the natural relationships in the data.

Estimation

Once the data have been collected, captured, coded, edited and imputed, the next step is estimation.

Estimation is the means by which the values for the population of interest are computed, so that meaningful conclusions can be drawn about that population, based on the information gathered from only a sample of the population itself.

For a sample survey for instance, the basis of estimation can be the unit´s weight, which indicates the average number of population units it represents. Then, a population total can be estimated, for instance by summing the weighted values of the sampled units, where the initial design weight is determined by the sampling design.

Sampling errors also occur in sample surveys, since only a portion of the population is enumerated and the sampled units do not have exactly the same characteristics of all the population units that

they represent. An estimate of the sampling error for each estimate should always be provided to indicate to end-users the quality of the resulting data, and also to better appreciate the relationships between the amount of resources allocated to the survey and the related outcomes.

Data analysis

Data analysis involves summarising the data and interpreting their meaning in a way that provides clear answers to the questions that motivated the survey. It should always relate the survey results to the issues identified by the Statement of Objectives. It is one of the most critical steps of a survey, since the quality of the analysis can substantially affect the usefulness of the whole survey.

Data analysis can be restricted to the survey data alone or it may compare the survey´s estimates with results obtained from other surveys or data sources. Often, it consists of examining tables and various summary measures, such as frequency distributions to summarize the data. Statistical inference can be used in order to verify hypotheses or study relationships between estimated attributes, for instance, using regression, chi-square tests or analysis of variance.

Documentation

Documentation provides a record of the survey and should encompass every survey step and phase. It may record different aspects of the survey and be aimed at different groups, such as management, technical staff and of course end-users.

Survey Design

Several steps have to be considered once the survey´s objectives are clarified: namely the target and survey population have to be identified.

Census vs Sample Survey

Two main kinds of surveys can be identified: census and sample surveys. A census collects information from all units of the population, while a sample survey collects information from only a fraction (typically small) of units of the population.

In both cases, the information collected is used to calculate statistics for the population as a whole, and often for subgroups of the population.

Several factors contribute to the choice between sample and census survey. The most influential are the following:

i. **Survey errors.** Commonly there are two types of survey errors: sampling and non-sampling errors. The sampling errors arise from estimating the population´s characteristics by inference from a small sample of the population itself. Non-sampling errors are all errors unrelated to

sampling. This might include, for instance, measurement and processing errors. Usually census surveys are considered much more accurate than sample survey, but the impact of non-sampling errors should always be considered. Moreover, often in the case of sample surveys, more sophisticated strategies can be put in place to minimize non-sampling errors, reducing the effort required by the data collection itself.

ii. **Cost**. Since data collection is one of the largest costs of a survey, it is self-evident that a census survey is considerably more expensive than a sample survey.

iii. **Timeliness**. Often the data must be gathered, processed, and the results disseminated (or fed into further processing stages) within a relatively short time-frame. This is even more important when the survey must capture attributes of a population that are changing with time. A census survey, which usually requires a much longer time to be carried out, could thus be inappropriate.

iv. **Size of population**. For small populations, a census survey might be preferable. This is because to produce accurate estimations, it might be necessary to sample a large fraction of the population. It could then be worth considering a small extra-cost to implement a census survey and have a complete characterization of the population. Conversely, when the population is large, a sample survey is usually preferable.

v. **Small area estimation**. Similar considerations as provided in the former point apply for small area estimation. This is particularly true if several sub-groups of the population have to be characterized.

vi. **Prevalence of attributes**. If the survey´s scope is to estimate the proportion of the population with a certain attribute, and this attribute is relatively frequent, a sample survey might be adequate. But if the considered attribute is infrequent, or rare, a census survey might be necessary.

    It is also true that in this case, particular sampling designs can be used to address this issue, such as stratified sampling. In case nothing is known about the relative frequency of the attributes to be surveyed, a small pilot survey could be conducted to obtain a preliminary evaluation of the population characteristics.

vii. **Specialized needs**. In some instances, the nature of the data collected requires highly trained personnel or expensive measuring equipment. In these cases, it might be impossible to conduct a census. This is often the case when the seismic vulnerability of a building has to be thoroughly evaluated. The task requires skilled engineers to collect and evaluate observations, measures and samples. This process cannot be repeated for the hundreds to thousands of structures which compose a mid-size town, and a more efficient approach must therefore be devised.

viii. **Other factors**. There could be specific reasons why to conduct a census survey. One would be to create a survey frame, which could be then used for subsequent sample surveys which use the same population.

    Another reason to conduct a census survey is to obtain benchmark information. This information can then be used to validate, cross check or improve the estimates from future sample surveys.

Target and Survey Population

When formulating the Statement of Objectives for a survey, the first concept to be defined is the target population, which is the population about which the information is desired. Usually a conceptual target population is first defined, describing the features of the items composing the population and the attributes of interest. Based on the analysis of the target population, a survey population is defined, which is the population actually covered by the survey.

Target and Survey populations should be very similar, but not necessarily identical. For instance, the cost or the logistical difficulties of collecting data in remote or isolated areas may lead to the exclusion of several units from the target population.

Survey frame

The survey frame, or sample frame, provides the means of identifying the elements of the survey population. The sample frame therefore ultimately defines the survey population.

The survey frame should include some or all of the following items:

i. **Identification data**. These are the items of the frame that uniquely identify each sampling unit.

ii. **Classification data**. Classification data are useful for sample selection and for estimation. Classification data may also include a size measure to be used in sampling. Geographical classification is also useful, defining the reference boundary (e.g., province, census tract or CRESTA code[8]).

iii. **Maintenance data**. Maintenance data are required if the survey is to be repeated at another time. This includes, for instance, the vintage of the data and its history (dates of subsequent changes in the underlying data).

iv. **Linkage data**. These data are used to link the units of the survey frame to other data sources, in order to identify the same units across different operational environments. The linkage data can be used to input, update and cross-validate the survey data. For instance, in a survey about residential buildings, the corresponding identity code from the cadastral archives could be added to the survey frame.

There are two main types of frames: list and area frames.

List frames

A list frame can be defined as a conceptual or physical list of all units in the survey population. A conceptual list frame often refers to a population which is previously unknown, but comes into existence when the survey is being conducted. A physical list frame is an actual list, compiled from different existing sources (often administrative data, such as, for instance, cadastral archives).

[8] https://www.cresta.org/

Area frames

An area frame is a special kind of list frame where the units on the frame are geographical areas. The survey population is located within these geographical areas.

Area frames can be used when the survey is geographical in nature (such as retrieving the attributes of the buildings of a town) or when an adequate list frame is unavailable, in which case the area frame can be used as a way for creating a list frame.

As a matter of fact, a list frame is often inadequate: populations can change over time (units can be created or destroyed, or change) therefore making a list frame out of date. Conversely, the geographical boundaries which define the area frame are more stable.

Area frames are often made up of a hierarchy of geographical units. Frame units at one level can be subdivided to form the units at the next level. Large geographical areas like provinces may be composed of districts or municipalities with each of these further divided into smaller areas, such as city blocks. In the smallest sampled geographical areas, the population may be listed in order to sample units within this area.

Sampling from an area frame is often performed in several stages. For example, suppose that a survey requires that dwellings in a particular town be sampled, but there is no up-to-date list. An area frame could be used to create an up-to-date list of dwellings as follows: at the first stage of sampling, geographical areas are sampled, for example city blocks. Then, for each selected city block, a list frame is built by listing all the dwellings of the sampled city block. At the second stage of sampling, a sample of dwellings is then selected. One benefit of such an approach is that it keeps the cost of creating the survey frame within reasonable bounds and it concentrates the sample in a limited number of geographical areas, making it a cost-effective way of carrying out challenging surveys. In the described example the overall frame is called a *multiple frame*.

The quality of a frame should be assessed based on the following criteria:

i. **Relevance**. Relevance should be considered as the extent to which the frame corresponds and permits accessibility to the target population. The more it differs from the target population, the larger the difference between the survey and target populations. Also, the utility of the frame for other surveys covering the same target population is a critical measure of its relevance.

ii. **Accuracy**. The accuracy of the data on the frame greatly affects the quality of the survey´s output. Different characteristics should be considered: coverage errors should be evaluated (under coverage and over coverage). Then, classification errors should be considered: for example are all units (properly) classified?

iii. **Timeliness**. Timeliness should be measured in terms of how up-to-date the frame is with respect to the survey´s reference period. If the information on the frame is substantially out-of-date (for instance due to the length of time necessary to build the frame itself), then some measures have to be implemented in order to improve timeliness.

iv. **Cost**. Several cost components can be considered. First, the total cost incurred to construct the frame should be determined. Second, the cost of the frame should be compared with the total survey cost. Third, the cost of maintaining the frame should be considered. To improve cost-effectiveness, a frame should be used by several surveys.

Survey errors

Several problems can arise during a survey that, if not anticipated and controlled, can introduce significant errors into the data. Therefore, every effort should be made in planning the design and implementation of the survey. Survey errors come from a variety of different sources. Two main categories can be defined: sampling errors and non-sampling errors.

Sampling errors

Sampling means estimating a population's characteristics by measuring only a small portion of the population. This process is naturally subject to errors, usually quantified by the sampling variance. The sampling variance measures the extent to which the estimate of a characteristic from different possible samples of the same size and same design differ from one another.

The estimate´s estimated sampling variance must be compared with the size of the survey estimate: if the variance is relatively large, then the estimate is unreliable and imprecise.

Several factors affect the magnitude of the sampling variance, including:

- The natural variability of the characteristic of interest in the population. The more variable the characteristic of interest in the population, the larger the sampling variance.
- The size of the population. This has usually an impact on the sampling variance only for small to moderate sized population.
- The sampling design and the method of estimation. For the same sample size and method of estimation, some sampling designs are more efficient than others, thus leading to a smaller sample variance.

Nonsampling errors

Aside from sampling error, a survey is also subject to a wide variety of errors not related to the process of sampling. These errors are commonly called nonsampling errors. They are present in both sample surveys and censuses (unlike sampling errors, which are only present in sample surveys). Nonsampling errors can be classified into two groups:

- **Random errors.** Random errors approximately cancel out over a large enough sample size, but also lead to increased variability.
- **Systematic errors.** These errors tend to accumulate over the sample leading to a bias in the final results. This bias is not reduced by increasing the size of the sample, and is responsible from most of the problems in the quality of a survey.

Sampling design

There are two types of sampling: non-probability and probability sampling. The one chosen depends primarily on whether reliable inferences are to be made about the population.

Non-probability sampling refers to a subjective method of selecting units from the population. It is relatively fast and easy compared with the probability sampling, but it strongly relies on the assumption (sometimes risky) that the sample is representative of the population.

Probability sampling on the other side involves the selection of units from the population based on a randomized selection. With respect to the non-probability sampling, more reliable estimates can be produced along with estimates of the sampling error, thus allowing inferences about the population to be made.

Non-probability sampling

Non-probability sampling is a method of selecting units from the population using a subjective method. This method does not require a complete survey frame, and is an easy and inexpensive way to obtain data. The problem is that it is often unclear whether it is possible to generalise the results from the sample of the population.

The non-probability sampling can be used for preliminary studies and pilot surveys, as a simple exploratory analysis of a previously unknown population. Several methods of non-probability sampling exist. In this document, we will mention only two of them, among the most used in risk-assessment applications: *volunteer sampling* and *quota sampling*.

Volunteer sampling

Volunteer sampling is based on the efforts of people volunteering to provide (or to collect) data. An example of volunteer sampling is OpenStreetMap (OSM), where geographical information is collected by an interacting community of users. Every OSM user decides autonomously what and where to conduct its mapping activities. This results in a large amount of data being collected with a significant selection bias. Nevertheless, this kind of information can be useful if particular care is used in any kind of inferences based on the collected data.

Quota sampling

With this method, sampling is done until a specific number of units (quota) for various subpopulations are selected. Quota sampling is a means to satisfy sample size objectives for the subpopulations.

Quota sampling is similar to stratified sampling in that similar units are grouped together, but it differs in how the units are selected. In probability sampling, the units are randomly selected, while in quota sampling a non-random method is used. As with all other non-probability sampling, in

order to make inferences about the population, it is necessary to assume that the units selected are similar to those not selected. Such assumptions are rarely verified.

Probability sampling

Probability sampling allows inferences to be made about the population based on observations from a sample. The sample should therefore not be subjected to selection bias.

There are many different types of probability sampling designs. The most basic is simple random sampling, and the designs increase in complexity to encompass systematic sampling, probability-proportional-to-size sampling, cluster sampling, stratified sampling and multi-stage sampling (other existing methods such as replicated sampling and multi-phase sampling will not be discussed here). Each of these techniques is useful in different situations. For instance, if the objective of the survey is simply to provide overall population estimates and stratification would be inappropriate or impossible, then simple random sampling would be the best choice. If the cost of the data collection in the survey is high with respect to the available resources, then cluster sampling may be used (if the characteristics of interest are not expected to be homogeneous within the clusters). If subpopulation (or geographically nested) estimates are also desired, then stratified sampling is usually performed.

In the following, most of the sampling designs will refer to an area frame, the one most used to generate and manage geo-information within the framework of the assessment of natural risks.

Simple random sampling (SRS)

The starting point for all probability sampling design is simple random sampling (SRS). SRS is a one-step selection method that ensures that every possible sample of size n has equal chances of being selected. As a consequence, each unit in the sample has the same inclusion probability. This probability, π, is equal to n/N where N is the number of units in the population.

Sampling with SRS can be done with or without replacement. Sampling with replacement allows for a unit to be selected more than once. Sampling without replacement means that once a unit has been selected, it cannot be selected again. SRS with or without replacement are practically identical if the sample is small compared with the size of the population. Generally, sampling without replacement yields more precise results and is operationally more convenient.

SRS has a number of advantages:

- It is the simplest sampling technique.
- It requires no additional (auxiliary) information about the frame (only a complete list of the survey population).

and disadvantages:

- It makes no use of auxiliary information when this information exists in the survey frame, even when using auxiliary information would improve the statistical efficiency.
- It can be expensive, since the sample may be spread out geographically.

- It is possible to draw “bad” samples that are not well dispersed and poorly represent the population.

Systematic sampling (SYS)

In systematic sampling (SYS), units are selected from the population at regular intervals. SYS can be used when no list is available, or when the list is roughly randomly ordered. Only a sampling interval and a random start are required. If no list of the population units is available in advance, a conceptual frame can be constructed by sampling every k unit until the end of the population is reached.

In the case of an area frame, systematic sampling refers to sampling from locations that are regularly arranged in the area frame (the interval is therefore a fixed displacement vector).

One problem with SYS is that the sample size is not known until the sample has been completely selected. Another problem can arise when the sampling interval (or displacement vector in case of an area frame) matches some periodicity in the population. In this case, an uncontrollable bias can be introduced into the collected data.

Probability-proportioned-to-size sampling (PPS)

Probability-proportional-to-size (PPS) sampling is a technique that uses auxiliary data and yields an unequal probability of inclusion. If population units vary in size (or another parameter which is expected to be correlated with the characteristic of interest of the survey), such information can be used during sampling to increase the statistical efficiency.

The main advantage of PPS sampling is that it can improve the statistical efficiency of the sampling by exploiting auxiliary information. This can result in a dramatic reduction of the sampling variance compared with SRS (or even stratified sampling page 30).

The disadvantages of PPS sampling are:

- It requires a survey frame that contains good quality, up-to-date auxiliary information for all units within the frame that can be used as size measures.
- It is inappropriate if the size measures are not accurate or stable. In such a case, it is better to use a cluster sampling or a stratified sampling.
- It can result in a sampling strategy that is less efficient than SRS if the survey variables are not correlated with the size variables.
- The estimation of the sampling variance is more complicated.
- The frame creation is more complex and costly with respect to SRS or SYS, since the size, (or the other parameter directly related to the targets of the survey) of each unit of the population has to be stored.

Cluster sampling

Cluster sampling is the process of randomly selecting complete groups (clusters) of population units from the survey frame. It is usually a less statistically efficient sampling strategy with respect to SRS and is used for several reasons:

- Cluster sampling can greatly reduce the cost of data collection, particularly if the population is spread out geographically.
- Sampling units from the population is not always practical. Sometimes, sampling groups of population units is much easier.
- It allows the production of estimates for the clusters themselves.

Cluster sampling is a two-step process. First, the population is grouped into clusters (this may consist of natural clustering, e.g., households, schools, etc.). The second step is to select a sample of clusters and select all units within the selected clusters.

In order for estimates to be statistically efficient, the units within a cluster should be as different as possible. Otherwise, if a certain degree of homogeneity (or spatial correlation) exists, the resulting data could be biased. This unfortunately is often the case, especially with geographical area-based sampling (see Tobler´s Law).

Stratified sampling

With stratified sampling (STR), the population is divided into homogeneous, mutually exclusive groups called strata, and then independent samples are selected from each stratum. Any of the sampling designs mentioned in the preceding sections can be used to sample within strata. For instance, with cluster sampling, it is common to first stratify, then draw the cluster sample (this is called stratified cluster sampling).

A population can be stratified by any variables that are available for all units of the frame, prior to the survey being conducted. Commonly used stratification variables include geographical information (political boundaries, or other spatial segmentations) or, for list frames, household size, age, type of economic activity, etc..

There are three main reasons for stratification. The first is to make the sampling strategy more efficient than SRS. The second is to ensure adequate sample sizes for specific domains of interest for which the analysis is to be performed. The third is to avoid drawing “bad” samples.

For a given sample size and estimator, stratification may lead to lower sampling errors or, conversely, for a given sampling error, to a smaller sample size.

In order to improve the statistical efficiency of a sampling strategy with respect to SRS, there must be a strong homogeneity within the strata (i.e., units within the stratum should be similar with respect to the variable of interest) and the strata themselves must be as different as possible (with respect to the same variable of interest). Generally, this is achieved if the stratification variables are correlated with the survey variable of interest.

Stratification is particularly important in the case of skewed populations (i.e., where the distribution of values of a variable is not symmetric but *leans* towards the left or to the right).

In such cases, a few population units can exert a great influence on the estimates, increasing its sampling variance. Therefore, such units should be placed in a stratum by themselves to ensure that they do not represent other units in the population.

The second reason for stratification is to ensure adequate sample sizes for known domains of interest, that is, subgroups of the population for which a separate estimate is required. For instance, in order to estimate the total number of buildings in a town, and at the same time the number of buildings in the districts of the town itself, a spatial stratification based on the boundaries of the districts can be conducted.

The third reason for stratifying is to protect against drawing a “bad” sample. In the case of SRS, the selection of the sample is completely random, therefore, especially for small samples, an apparent clustering (or other spatial inhomogeneity) of the sampling points can be observed. If, moreover, a single subgroup of the population has a smaller probability of inclusion, a SRS sampling might under sample it.

Inclusion probabilities usually vary from stratum to stratum; it depends on how many samples are allocated to each stratum. To calculate the inclusion probability for most sampling designs, the size of the sample and the size of the population in each stratum must be considered.

The advantages of stratified sampling are:

i. It can increase the precision of overall population estimates, resulting in a more efficient sampling strategy. A smaller sample can save a considerable amount of effort being expended on the survey, hence decreasing the cost of the data collection.

ii. It can guarantee that important subgroups, when defined as strata, are well represented in the sample, resulting in statistically efficient domain estimators.

iii. It can protect against selecting a “bad” sample.

iv. It allows different sample frames and procedures to be applied to different strata (e.g., SRS in one stratum, PPS in another).

The disadvantages of stratified sampling are:

i. It requires that the sampling frame contains high-quality auxiliary information for all units on the frame, not just those in the sample that can be used for stratification.

ii. It can result in a sampling strategy that is less statistically efficient than SRS for survey variables that are not correlated to the stratification variables.

iii. The estimation is slightly more complex than for SRS and SYS.

Multi-stage sampling

Multi-stage sampling is the process of selecting a sample in two or more successive stages. The units selected at the first stage are called primary sampling units (PSU´s), units selected at the second stage are called secondary stage units (SSU´s), etc. The units are different at each stage, differing in structure and are hierarchical. In two-stage sampling, the SSU´s are often the units of the population.

Multi-stage sampling is commonly used with area frames to overcome the inefficiencies of one-stage cluster sampling (if the neighbouring units in a cluster are similar, then it is more statistically efficient to sample a few SSU´s from many PSU´s). Multi-stage sampling usually have two to three stages, in order not to increase the complexity of the estimation.

The advantages of multi-stage sampling are:

i. It can result in a more statistically efficient (i.e., a reduction of the sample size) sampling strategy than a one-stage cluster design when clusters are homogeneous with respect to the variable of interest.

ii. It can greatly reduce the travel time and cost of data collection as a result of the sample being less dispersed then for other forms of sampling such as SRS.

iii. It is not necessary to have a list frame for the entire population. All that is needed is a good frame at each stage of the sample selection.

The disadvantages of multi-stage sampling are:

i. It is usually not as strategically efficient as SRS.

ii. The final sample size is not always known in advance, since it is not usually known how many units are within a cluster until after the survey has been conducted (the sample size can be controlled, however, if a fixed number of units are selected per cluster).

Multi-phase sampling

Despite the similarities in name, multi-phase sampling is quite different from multi-stage sampling. Although multi-phase sampling also involves taking two or more samples, all samples are drawn from the same frame and the units have the same structure at each phase. A multi-phase sample collects basic information from a large sample of units and then, for a subsample of these units, collects more detailed information.

Multi-phase sampling is useful when the frame lacks auxiliary information that could be used to stratify the population or to screen out part of the population.

Multi-phase sampling can also be used to collect more detailed information from a subsample when there is insufficient budget to collect information from the whole sample. Similarly, multi-phase sampling can be used when there are very different costs of collection for certain characteristics of the population. Consider a survey aiming at compiling an exposure and vulnerability model of a town. While basic features of the building stock can be rapidly collected in the inspected buildings (considered units of the population), a thorough assessment of the vulnerability requires a detailed inspection of the building by expert personnel, hence it is by comparison much more expensive. The survey could be done as a two-phase sample, with the basic features collected at the first phase, and only the second, smaller survey focusing on the vulnerability assessment.

The advantages of multi-phase sampling are:

i. It can greatly increase the precision of estimates (compared with SRS).

ii. It can be used to obtain auxiliary information that is not in the sampling frame (in particular, stratification information for second phase sampling).

iii. It can be used when the cost of collection for some of the survey variables is particularly expensive or burdensome.

The disadvantages of multi-phase sampling are:

i. It takes longer to get results than from a one-phase survey, if the results of the first phase are required to conduct the second phase.

ii. It can be more expensive than a one-phase survey since it requires interviewing a sampled unit mode than once.

iii. If the characteristics of interest of the population change frequently, time delays between phases may pose problems.

Estimation

Estimation is the means by which different values for the population of interest can be obtained, such that specific conclusions about the population can be drawn based on information gathered from only a sample of the population.

The estimation process in a probability survey is based on the assumption that each sample unit represents not only itself, but also several units of the survey population. The average number of units in the population that a sample represents is called the design weight of the unit. Determining the weight is an important part of the estimation process.

Once the final estimation weights have been calculated, they are applied to the sample data in order to compute estimates. Typical estimates for the survey population are *totals, averages and proportions*, which can be computed for a wide range of characteristics collected from the sample units.

The design weight therefore can be defined as the average number of units in the survey population that each sampled unit represents, and is determined by the sampling design. The design weight w_d for a unit of the sample is the inverse of its inclusion probability, π, that is the probability of the unit to be selected in the sample.

For a multi-stage or multi-phase sampling design, a unit´s probability of selection is the combined probability of selection at each stage or phase. For instance, for a two-phase sample where a unit´s probability of selection is π_1 at the first phase and π_2 at the second phase, a sample unit´s design weight is:

$$w_d = \frac{1}{\pi_1} \cdot \frac{1}{\pi_2}$$

The total number of units in the survey population is therefore calculated by summing the weights[9]:

[9]We are not considering non-response, therefore assuming that for each unit in the sample, a corresponding estimate can be conducted. In cases where this assumption is not reasonable, the weights have to be adjusted accounting for the non-response rate.

$$\hat{N} = \sum_{i \in S_r} w_i \quad (1)$$

$$= \sum_{i \in S_r} \frac{1}{\pi_i} \quad (2)$$

For quantitative data, the estimate of a total value is the product of the final weight w_i and the value y_d for each sampled unit summed over the sample:

$$\hat{Y} = \sum_{i \in S_r} w_i y_i \quad (3)$$

In a similar way the population average is defined as:

$$\hat{\bar{Y}} = \frac{\sum_{i \in S_r} w_i y_i}{\sum_{i \in S_r} w_i} \quad (4)$$

$$= \frac{\sum_{i \in S_r} w_i y_i}{\hat{N}} \quad (5)$$