

Assessment of Seismic Building Vulnerability from Space

Christian Geiß,^{a) b)} Hannes Taubenböck,^{a)} Sergey Tyagunov,^{c)} Anita Tisch,^{d)}
Joachim Post,^{a)} Tobia Lakes^{b)}

This paper quantitatively evaluates the suitability of multi-sensor remote sensing to assess the seismic vulnerability of buildings for the example of the city Padang, Indonesia. Features are derived from remote sensing data to characterize the urban environment and are subsequently combined with *in situ* observations. Machine learning approaches are deployed in a sequential way to identify meaningful sets of features that are suitable to predict seismic vulnerability levels of buildings. When assessing the vulnerability level according to a scoring method, the overall mean absolute percentage error is 10.6%, if using a supervised Support Vector Regression approach. When predicting EMS-98 classes, the results show an overall accuracy of 65.4% and a *Kappa* statistic of 0.36, if using a naïve Bayes learning scheme. This study shows potential for a rapid screening assessment of large areas which should be further explored in the future.

INTRODUCTION

Most casualties from earthquakes are associated with collapsing buildings. Therefore the rapid urbanization observed in earthquake prone regions places more people at risk than ever before. As a consequence, the death toll in urban areas is expected to reach unprecedented levels (Bilham 2009). Developing countries are characterized by dynamic urban growth with large shares of unplanned, spontaneous and often highly vulnerable settlements. Simultaneously, these settlements are highly variable over short time scales (Wieland et al.

^{a)} German Aerospace Center (DLR), German Remote Sensing Data Center (DFD), Oberpfaffenhofen, Germany

^{b)} Humboldt Universität zu Berlin, Geography Department, Berlin, Germany

^{c)} German Research Centre for Geosciences (GFZ), Potsdam, Germany

^{d)} Institute for Employment Research (IAB), Nuremberg, Germany

2012). In this regard the continuous assessment and monitoring of the seismic vulnerability of buildings is a challenging task, especially when large-area evaluations are required. Numerous studies emphasize that remote sensing can play a valuable role in supporting the extraction of relevant features for pre-event vulnerability analysis of built-up structures (French and Muthukumar 2006, Mueller et al. 2006, Sarabandi et al. 2008, Taubenböck et al. 2009a, Borfecchia et al. 2010, Sahar et al. 2010, Borzi et al. 2011, Deichmann et al. 2011, Wieland et al. 2012, and GEM 2013). The intrinsic advantage of remote sensing is the ability to offer an overview of building stocks and serve as a screening method for derivation of building vulnerability related features, such as shape characteristics, height, roof material, year of construction, structure type and spatial context (Geiß and Taubenböck 2013).

Approaches evolved from engineering science to assess the seismic vulnerability of buildings such as the quantification of displacement response spectra (Crowley et al. 2004), capacity spectrum (Freeman 2004), or damage probability matrices (e.g. EMS-98) (Grünthal et al. 1998) consider single structures, studied in a detailed analytical way, or aggregated and idealized classes of buildings, that can be assessed for large-area applications (Calvi et al. 2006). Existing studies have focused on two different approaches linking remote sensing data to seismic building vulnerability: (i) by defining a direct relation, by for example, vulnerability curves based on features that can be gained from remote sensing data (Taubenböck et al. 2009a; Borzi et al. 2011) or (ii) by using remotely sensed data primarily for spatial inter- and extrapolation of *in situ* surveys, by for example, supervised classification techniques (Borfecchia et al. 2010). The former approach can only perform well if finding and defining valid vulnerability curves based on the available input features for respective built-up structures (Taubenböck et al. 2009a). Analogously, the accuracy of the latter approach is highly depending on finding significant and robust proxy variables that have a high correlation with *in situ* observations. Typically only a small number of features that influence seismic building vulnerability, such as building height or shape features are reflected directly or can be gained from remote sensing data. Naturally, both approaches are further dependent on the accuracy of the derived features and the precision of the results may vary with respect to local idiosyncrasies and different built-up structures. To date, studies have evaluated the potential of remote sensing in a solely qualitative manner (Mueller et al. 2006) or presented results that emphasize the viability of the use of remote sensing (Borfecchia et al. 2010), but lack the identification and documentation of the necessary and meaningful features. The goal of this paper comprises the quantitative evaluation of the

potential and limitations of satellite remote sensing for assessing seismic building vulnerability in Padang, Indonesia. High resolution optical imagery, height information from a normalized digital surface model (nDSM), and multi-temporal medium resolution optical data, are used to calculate features that characterize the urban environment. These are combined with *in situ* data about the buildings' vulnerability levels. The *in situ* assessed vulnerability levels are based on a scoring method and the well-established EMS-98 scheme.

By using techniques of machine learning feature selection we aim to identify the most relevant features that can be used for the estimation of seismic building vulnerability. Techniques of regression analysis are also utilized to determine the strength, direction and significance of independent variables on the dependent variable, namely *in situ* assessed vulnerability level. Therefore, our first research question is as follows:

1. Which features can be derived from satellite remote sensing data that best explain seismic building vulnerability?

Subsequently, we assess the accuracy of supervised regression and classification techniques to answer the question:

2. How suitable are features derived from satellite remote sensing data for estimating seismic building vulnerability levels?

EXPERIMENT SETTING AND DATA

STUDY AREA AND GENERAL EXPERIMENT SCHEME

We chose the city of Padang, Indonesia because it is situated in one of the most earthquake-prone regions in developing countries worldwide. Padang is located in West Sumatra and is the capital city of the Sumatera Barat province. It has approximately one million inhabitants and represents the third largest city on the island of Sumatra (Figure 1). The dynamic urban system of Padang is characterized by a high concentration of inhabitants and infrastructure. Padang has supra-regional relevance with an international airport and railway connection. The city plays an essential economical role for the coastal region and mountainous hinterland. Off the coast of Padang the Sunda Arc marks an active convergent plate boundary, placing the city in a zone of extremely high probability of severe earthquakes and secondary effects such as tsunamis (Taubenböck et al. 2009b). On September 30th 2009, Padang was hit by an earthquake with a moment magnitude of $M_w = 7.6$. Despite its size, the

Sunda megathrust was not ruptured and the stress on the Mentawai segment, which was accumulated over 200 years, was not significantly reduced. The megathrust strain-energy budget remains at a high level, threatening a great, and also tsunamigenic earthquake with a magnitude $M_w > 8.5$ on the Mentawai patch (McCloskey et al. 2010).

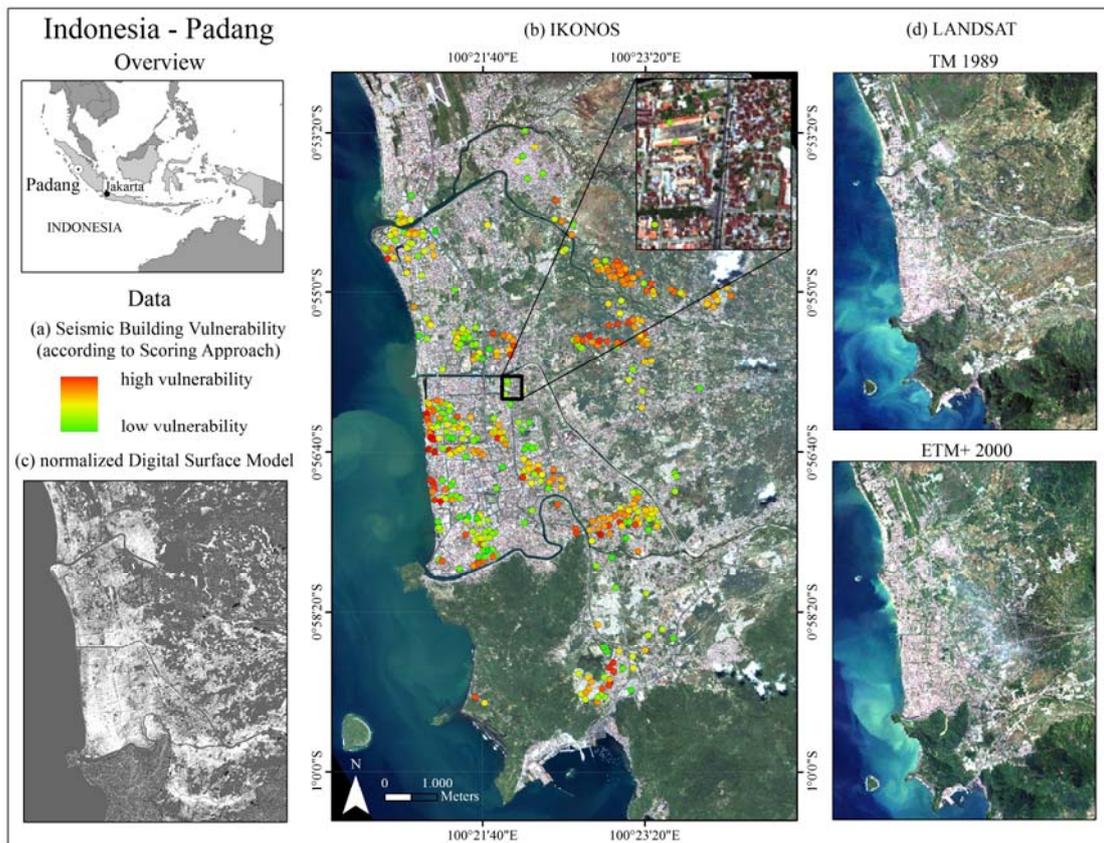


Figure 1. Overview on the location of the study area Padang, Indonesia, *in situ* and remote sensing data; (a) *in situ* assessed buildings represented as points and superimposed on (b) multispectral IKONOS imagery; (c) normalized digital surface model (nDSM) as basis for height estimations; (d) multitemporal LANDSAT data for spatiotemporal analyses.

To evaluate the potential of remote sensing for assessing the seismic vulnerability of Padang's buildings we followed the schematic workflow shown in figure 2. Based on remote sensing data we calculated features on two different spatial levels, building and block level (see Table 2 for a detailed list of features). Subsequently, the *in situ* data with affiliated vulnerability information was added. Techniques of machine learning based feature selection and regression analyses were then used to identify features that were most suitable to assess seismic vulnerability levels and quantify their explanatory content. Under consideration of

the results, we built several supervised regression and classification models and assessed the accuracies of the predictions. In-depth explanations are given throughout sections “experiment setting and data” and “methods”.

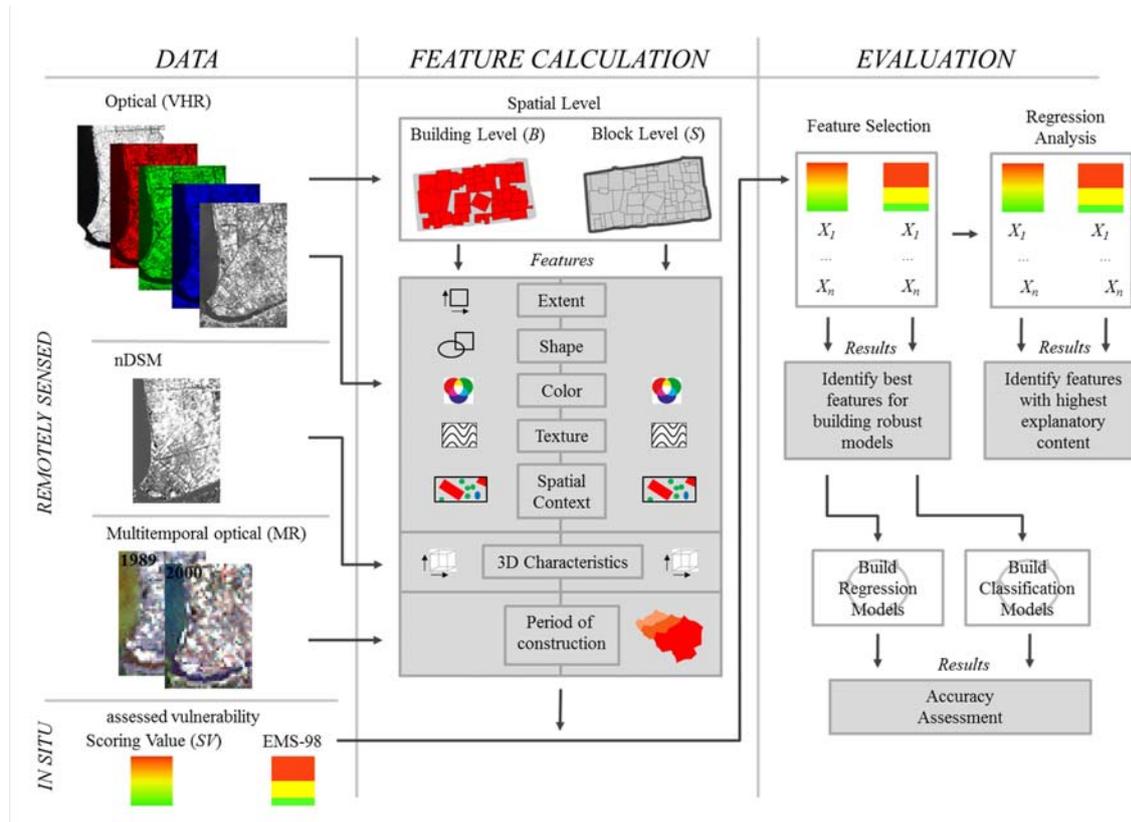


Figure 2. Overview of the general experiment scheme followed in this study, which is subdivided according to the main categories “data”, “feature calculation” and “evaluation”. Detailed explanations are given throughout sections “experiment setting and data” and “methods”.

IN SITU DATA AND SEISMIC VULNERABILITY ASSESSMENT

The *in situ* data on seismic vulnerability in Padang was collected in February/March, 2008 within the “Last-Mile” project (Taubenböck et al. 2009b). The building inventory database compiled for Padang, based on a ground truth survey, includes information about physical characteristics of 434 buildings in the city. The sampling scheme of the buildings aimed at both the incorporation of all existing housing types of Padang (from informal small shacks to high-rise commercial buildings) and broad spatial coverage (from the urban center to periphery suburbs) (Figure 1). The database includes information about geometry, material

of bearing structures and walls, foundations and local soil conditions, material of the roof, type of building, etc. For most of the inspected buildings the dataset is supplemented by results of physical tests providing information about the reinforcement and quality of concrete of the main bearing structures. Additionally, for every building in the database there is an indication of the damage level due to previous earthquakes (Taubenböck et al. 2013). Based on this data, a vulnerability scoring approach and a classification according to the EMS-98 scale was carried out to assess the seismic vulnerability of the buildings as explained in the next two subsections. We incorporate both methods in this study to provide an assessment that reflects local idiosyncrasies and expert knowledge (scoring approach), and an assessment according to a more generalizable and wide-spread scheme (EMS-98). The working definition of seismic building vulnerability of this paper is related to engineering driven definitions, whereby vulnerability describes the probability of damage to a building under specified earthquake influence (Whitman 1973).

DATA FROM SCORING APPROACH

We incorporate data from a vulnerability scoring approach, which was carried out after the *in situ* data collection within the “*Last-Mile*” project. It is based on an indexing method (see e.g. Calvi et al. 2006 for a general description of such methods) that incorporates local expert knowledge. The aforementioned documented building parameters are first expressed quantitatively on a normalized scale. High values express characteristics that are considered as favorable, regarding the buildings’ seismic vulnerability. Subsequently, individual weights for the respective parameters are assigned based on expert observations made in the study area during previous research. For instance, an established method to assess the stability properties of reinforced concrete is the aforementioned Schmidt rebound hammer test. Completed structural survey results in the study area confirmed the hammer test outcomes to be the most important indicator and are therefore given the highest weight. Finally the weighted values of the respective parameters are summed up. For brevity the reader is referred to Mück et al. (2013) for a detailed description of conceptual and methodological details of this approach. Figure 3 shows the distribution of the scoring values within an interval of [10.85, 25.6], a mean value of 18.4 and a standard deviation of 2.89. Lower scoring values express higher building vulnerability, while higher values represent lower vulnerability.

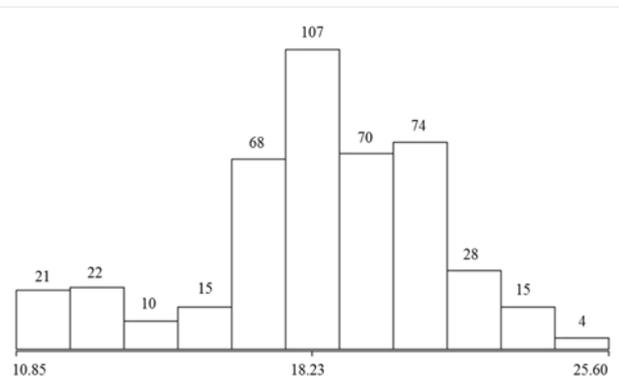


Figure 3. Distribution of the scoring values for 434 *in situ* surveyed buildings within an interval of [10.85, 25.6]

EMS-98 CLASSIFICATION DATA

The classification scheme of the European Macroseismic Scale (EMS-98) (Grünthal et al. 1998) was initially designed for Europe. However, it is accepted as a useful point of reference for other areas when collapse data are absent (Jaiswal et al. 2011). According to the EMS-98 different types of buildings are classified into six vulnerability classes, denoted alphabetically from A (highest vulnerability) to F (lowest vulnerability). The classification depends primarily on the building material and the type of structure, taking into consideration a variety of additional factors (such as constructional and architectural features, quality of materials and workmanship, age and state of preservation, etc.), which may affect the seismic performance of the buildings. As mentioned, this kind of information is contained in the *in situ* data set. Additionally, available reports of the West Sumatra Earthquake of September 30th, 2009, containing descriptions of damages for different types of buildings in the affected area, are considered (Sengara et al. 2010). It is worth noting that among the main causes of damage reported, poor quality of materials and construction was mentioned. This should be taken into account when assessing the seismic vulnerability of existing buildings in the area.

As a first step, all buildings were differentiated based on the material of the main structural elements: steel, concrete, bamboo, and wood. For the case of steel structures, according to the EMS vulnerability table, the class E can be assigned as the most probable one with the range of probable classes C-F. For wooden structures the class D can be assigned with the range of B-E. Bamboo structures are not included in the EMS-98, however we considered that they are similar to, but slightly more vulnerable than wooden structures.

We therefore assumed a proper range for these structures is C-D (with the less probable range of B-E). In the case of concrete structures, we had to keep in mind the existing uncertainty of the structural type (in the database there is no indication if the structures are made of concrete, masonry units, or reinforced concrete frame). Therefore for concrete structures we assigned the class C initially, with the less probable range of A-E. Secondly, we took into account weaknesses that are mentioned in the inventory database. If necessary, the corresponding modifiers were applied to the initial vulnerability class. Assessment of weaknesses should be building-type-specific, by examining if the principal rules of earthquake-resistant design (including quality, regularity, homogeneity, ductility, overall integrity and stability of the structure) are observed. The following essential weakness characteristics were considered: wide openings of the ground floor (which may cause soft storey effects), very thin bearing elements (columns and beams) or their absence, lack of reinforcement (or low reinforcement ratio) of the main structural elements, poor results from the hammer tests. If some of the listed weaknesses can be identified, the vulnerability class of such buildings was downgraded. Besides weaknesses there may be strengths (e.g. relatively high reinforcement ratio or enlarged dimensions of the bearing structures), however, we do not take those into account due to the reported poor quality of materials and workmanship in the area in general. Another illustrative indicator of seismic performance of structures is the damage observed from previous earthquakes. Based on this we downgraded the vulnerability class in the case of moderate to severe damage to class B or even to the range of A-B, depending on the structural type. Table 1 shows the EMS-98 classes assigned and the affiliated number of buildings. For some classes there were only a sparse numbers of instances, and so some classes were aggregated. The bold letter indicates the most likely vulnerability class. This is done to have a more sufficient number of samples per class for applying supervised learning approaches.

Table 1. EMS-98 classes and affiliated number of instances

<i>Initial classification</i>	<i>Number of instances</i>	<i>Aggregated classification</i>	<i>Number of instances</i>
A-B	7	A-B	242
B	235		
B-C	26	B-C	149
C	123		
C-D	32	C-D-E	43
D-E	3		
E	8		

REMOTE SENSING DATA

As mentioned, no remote sensing data set or derived products offer the complete set of features that are frequently used for analytical engineering assessments. However, remote sensing offers a huge spectrum of sensor systems that may deliver useful data for a subset of seismic vulnerability indicators. Active remote sensing, such as airborne LiDAR (Light Detection And Ranging) measurements, are frequently used to extract buildings heights (e.g. [Sirmacek et al. 2012](#)). When combined with optical data, high resolution Synthetic Aperture Radar (SAR) can be utilized to derive crucial building features such as footprint and height (e.g. [Polli and Dell'Acqua 2011](#)). The latest generation of multispectral spaceborne sensors such as IKONOS, QuickBird-2, GeoEye-1 or WorldView-2 enables satellite based detection, characterization and assessment of objects in urban environments, on a large scale (e.g. [Maktav et al. 2005](#), [Weng and Quattrochi 2006](#), or [Rashed and Jürgens 2010](#)). Especially for developing countries and remote areas, the airborne acquisition of data is still a challenging task in terms of flight campaign preparation and keeping data costs reasonable. We therefore utilize data with resolution characteristics that can be achieved by spaceborne sensors for this study, which are comparatively low priced and easily available, although we are aware that especially airborne sensors, can achieve higher resolutions. For the evaluation of spaceborne remote sensing capabilities multispectral IKONOS imagery, height information from a nDSM and LANDSAT data were acquired within the “Last-Mile” project ([Taubenböck et al. 2009b](#)).

The optical IKONOS imagery (acquisition date: 2005-04-12) covers a spectral range of 0.445-0.853 μm , with a geometric resolution of 1 m in the panchromatic band and 4 m in the multispectral (blue, green, red, nir) bands. The satellite has a revisit capability of 1.5-3 days

and the sensor's swath-width of 11 km allows covering large parts of cities. The data were pansharpended and atmospherically corrected using the ATCOR (Atmospheric and Topographic Correction) model (Richter 1996). Height information was derived from a digital surface model (DSM) and a digital terrain model (DTM). Measurements from the models were based on the return signals received by two radar antennas mounted on an aircraft, and the application SAR interferometry techniques (see Li et al. 2004 for a detailed description of the technique applied). Both data sets have a geometric resolution of 5 m and a Root-mean-square-error (*RMSE*) regarding the vertical accuracy of 1 m. To receive relative height information of elevated objects a nDSM is calculated by subtracting the height values of the DTM from the height values of the DSM using map algebra. Although the height information for this study was acquired based on an airborne sensor, the data resolution characteristics can also be achieved or even refined when using along-track stereo data of spaceborne sensors, such as Cartosat-1, IKONOS, or WorldView-2 (Sirmacek et al. 2012). In addition, we used data from the LANDSAT sensors Thematic Mapper (acquisition date: 1989-07-25) and Enhanced Thematic Mapper (acquisition date: 2000-07-15). Both sensors have 7 multispectral bands covering a spectral range of 0.45-2.35 μm with a maximum geometric resolution of 30 m (TM) and 15 m (ETM+) (Irish 2008). The satellites have a revisit capability of 16 days and the data can be accessed free-of-charge within a public image archive which dates back to 1972.

METHODS

FEATURE CALCULATION

Before we calculated features from the remote sensing data we first manually digitized the building footprints and derived building blocks from a road network (Taubenböck et al. 2008a). In the subsequent steps, features were calculated for both. The building level was used for calculating features that refer to characteristics of the respective building (designated by the subscript *B*), whereas the block level (designated by the subscript *S*), was used to describe the spatial setting the respective buildings are embedded in. This is why we chose to use building blocks derived from a street network rather than artificial spatial units, such as quadratic objects. This allows us to reflect the urban morphology, which is constituted by distinct areas that are generally irregularly shaped. Simultaneously, the difficulty of having to determine the optimal kernel size *a priori* is avoided (Herold et al. 2003).

The calculated features relate to the building objects' two-dimensional extent and the description of their shape characteristics to directly reflect the influence of geometry on the vulnerability level. Through manual image interpretation a further distinction of buildings with "flat" and "non-flat" roofs was made. In addition, the multispectral information available was used to calculate both 1st and 2nd order statistical values on the building level as well as on the block level. Thereby, the spectral information on building level primarily serves as a descriptor of roof surface material and arrangement (Mueller et al. 2006); whereas the information on building block can be utilized to describe distinct urban structures (Herold et al. 2003, Wu et al. 2008). Mean and standard deviation values of the different image bands as well as band ratios, which are intended to emphasize spectral dissimilarities, were calculated. Rotation-invariant texture measures for the panchromatic and near-infrared band were calculated using both the co-occurrence matrix (GLCM, Haralick et al. 1973) and grey level difference vector (GLDV, Weszka et al. 1976). Texture is the term used to characterize tonal or gray-level variations in an image. It has been demonstrated that it can provide supplementary information to overcome a lack of spectral resolution (Pacifci et al. 2009).

Features explicitly describing the spatial context consist of the calculated distance to neighboring buildings, the area of building blocks and the average size of the buildings located within. Furthermore, spatial metrics such as proportion measures of land cover classes "buildings", "sealed", "grass/meadow", "trees", and combined classes such as "vegetation" and "impervious surface" are calculated based on an urban land cover map derived from the IKONOS data. The urban land cover map has a high overall accuracy of 97% correctly classified pixels based on an automated object based, hierarchical classification methodology and subsequent manual enhancement (Taubenböck et al. 2009b). Additionally, a semantic classification which is built on physical features that describe the urban morphology is incorporated. The classification describes the socio-economic status of the population by distinguishing "slums", "suburbs", "low income areas", "medium income areas", and "high income areas" (Taubenböck et al. 2009c). Housing clusters that describe the socioeconomic level of occupants represent a proxy variable that has already been used in previous studies for the assessment of seismic risk (Prasad et al. 2009). The incorporation of height information allows the calculation of 3D features such as building floor number, which was calculated with an accuracy of 86.7% (Taubenböck et al. 2009c), floor space, ratio of diameter and height as well the average building height within a building block. Slope

values are calculated at the block level to describe topographic location. By analyzing two Landsat images, the period of construction was derived at the individual building level based on a semi-automated post classification change detection procedure, with an overall classification accuracy of 89.1% for 1989 and 92.4% for 2000 (Taubenböck et al. 2008b).

Overall, each building object is represented by a 132-dimensional feature vector, whereby 73 features are related to individual buildings, and 59 provide block level information. The features calculated (i) aim to reflect features that went into the calculation of the *in situ* values when they can be quantified by means of remote sensing data, such as building height or geometry, (ii) represent the spectrum of features frequently utilized in previous studies on remote sensing based building vulnerability assessment, and (iii) features that were used to discriminate urban built-up structures by means of remote sensing previously.

Table 2. List of features derived from remote sensing data.

Sensor/ Data	Feature		Description
		Area _B	
	E	E	Perimeter _B
	X	N	Width _B
	T	T	Length _B
			Length/Width _B
			Asymmetry _B
			Elliptic Fit _B
I K O N O S			Radius of Smallest
			Enclosing Ellipse _B
			Radius of Largest
		S	Enclosing Ellipse _B
		H	Rectangular Fit _B
		A	Roundness _B
		P	Density _B
		E	Main Direction _B
			Border Index _B
			Shape Index _B
		Compactness _B	
		Roof Type _B	

Description of a building's geometry by approximate comparison of the shape with 2-dimensional geometrical forms such as square, rectangle or ellipse

Direction of a building's longer eigenvector

Description of how jagged/smooth a building's shape is

Discrimination of buildings with "flat" and "non-flat" roofs

		Max. Diff. B, S	Maximum difference of the image layers' mean intensity value
S P E C.	M	Brightness B, S	Mean brightness of the image layers
	A	Blue(1) B, S	Mean spectral value
	N	Green(2) B, S	
		Red(3) B, S	
		Nir(4) B, S	
I N F O	S	Blue(1) B, S	Standard deviation of the spectral values
	T	Green(2) B, S	
	D	Red(3) B, S	
	V.	Nir(4) B, S	
1 st O R D. S	B	$[M(4)-M(3)]/[M(4)+M(3)] B, S$	Mean spectral value (M) of respective band used for transformation according to formula
	A	Brightness $\times M(1)/M(2) B, S$	
	N	Brightness $\times M(1)/M(3) B, S$	
	D	Brightness $\times M(2)/M(3) B, S$	
		$M(1)/M(2) B, S$	
		$M(1)/M(3) B, S$	
		$M(1)/M(4) B, S$	
	R	$M(2)/M(3) B, S$	
	A	$M(2)/M(4) B, S$	
	T	$M(3)/M(4) B, S$	
	O	$M(2)/[M(1)+M(2)+M(3)+M(4)] B, S$	
	I	$[M(4)/M(3)]/[M(1)+M(2)+M(3)+M(4)] B, S$	
	S	$[M(3)-M(2)]/[M(3)+M(2)] B, S$	
	$[M(3)-M(1)]/[M(3)+M(1)] B, S$		
S P E C.		Angular 2 nd Moment $B \text{ pan nir, } S \text{ pan nir}$	Also called "Energy"; Quantification of texture uniformity by counting pixel pair repetitions
	G	Contrast $B \text{ pan nir, } S \text{ pan nir}$	Difference between the highest and the lowest values of a contiguous set of pixels
	L		Similar to "Contrast" but weighting the diagonal linearly and not exponentially
	C	Dissimilarity $B \text{ pan nir, } S \text{ pan nir}$	
	M	Std. Dev. $B \text{ pan nir, } S \text{ pan nir}$	Measures the dispersion of values around the mean
2 nd O		Mean $B \text{ pan nir, } S \text{ pan nir}$	Pixel value is weighted by the frequency of its occurrence in combination with a certain neighbor pixel value

R D.		Homogeneity _{B pan nir, S pan nir}	Describes homogeneity by measuring the presence of near diagonal elements in a GLCM
		Entropy _{B pan nir, S pan nir}	Quantifies the disorder by approximating the textural uniformity
		Correlation _{B pan nir, S pan nir}	Measures the linear dependency of pixel values
	G	Angular 2 nd Moment _{B pan nir, S pan nir}	
L D V		Contrast _{B pan nir, S pan nir}	Same as respective GLCM measure but based on a grey level difference vector
		Mean _{B pan nir, S pan nir}	
		Entropy _{B pan nir, S pan nir}	
S P A L T C O N T E X T		Distance to 1 st , 2 nd , 3 rd Neighbor Building _B	
		Area Building Block _s	
		Average Building Size _s	Average area of the buildings per building block
		Degree of Building Density _s	Percent of building block area covered by buildings
		Share of LC class "sealed" _s	Shares of respective land cover (LC) classes per building block
		Share of LC class "grass/meadow" _s	
		Share of LC class "tree(s)" _s	
		Share of Vegetation _s	Share of LC class "grass/meadow" + Share of LC class "tree(s)" _s
		Share of Impervious Surface _s	Degree of building density + Share of LC class "sealed" _s
		Structure Type _s	Classification of semantic structure types "Slum", "Suburb", "Low Income", "Medium Income", "High Income"
n D S M	3	Floor Number _B	Building height expressed in floors
	D	Floor Space _B	Area×Floor Number
		Diameter/Height _B	Length/Floor Number
	F U E R	Width/Height _B	Width/Floor Number
	A E T S	Average Building Height _s	Average building floor number per building block
	SLOPE	Slope _s	Average slope value per building block; Slope value per pixel corresponds to avg. degrees in 3 x 3 window

L A N D S A T	TM ETM+	Period of construction _s	Semi-automated classification of "urban"/"non-urban" areas for 1989 and 2000
---------------------------------	------------	--	--

Subscripts: B = Building Level, S = Block Level, Pan = Panchromatic, Nir = Near-Infrared, M = Mean Value;

FEATURE SELECTION ANALYSIS

The selection of features to be used for regression and classification models is generally a difficult task, especially when dealing with a huge number of features as in this study. These often exhibit redundancy, are highly correlated, and suffer from the “Hughes phenomenon”. The latter describes the effect that for a limited amount of samples the predictive power decreases as the dimensionality of the feature vector increases (Hughes 1968). Therefore, two machine learning based feature selection algorithms were applied on the data set. The feature selection approaches were chosen because they can handle both regression problems and evaluate discrete valued variables. Furthermore, one can discriminate algorithms which evaluate individual features and those which assess subsets of features (Hall & Holmes 2003). As such, the Relief-F (Kononenko 1994) approach was chosen because it enables to rank individual features. Additionally the Correlation-based Feature Selection (CFS) method was chosen, since it enables the scoring of the value of groups of features (Liu et al. 2002).

RELIEF-F

The focus of the Relief approach is to rank features according to how well their values enable the discrimination of cases that are near to each other. The underlying assumption is that a suitable feature should have different values for cases from different classes, and similar values for cases from the same class (Liu and Schumann 2005). This principle is implemented by random sampling of an instance from the input data and the subsequent locating of its nearest neighbor from the same and opposite class. Values of the features of the nearest neighbors are then compared to the sampled instance and used to update relevance values for each feature. We used the enhanced approach “Relief-F” presented by Kononenko (1994), which allows handling of multi-class data (Robnik-Šikonja and Kononenko 2003).

Relief-F also smoothes the influence of noise in the data by averaging the contribution of k -nearest neighbors of each sampled case (Hall and Holmes 2003).

Primarily two parameters need to be defined when calculating relevance values: the number of cases m and the number of nearest neighbors k . According to Liu and Schumann (2005), m is also set to be the number of all training cases since a larger number of cases implies more reliable approximation, and k is set to 10, without weighting the nearest neighbors according to their distance. We tested several values for k (10, 20, 30, 40), the results of which showed little sensitivity to k values in this study, with only very moderate variations of the relevance value $W(A)$. $W(A)$ gives an indication in the interval $[-1, 1]$ to what degree the respective feature is relevant. If $W(A) > 0$ then there is some degree of relevance, whereas features with a value $W(A) < 0$ are not relevant.

CORRELATION-BASED FEATURE SELECTION (CFS)

Under the premise that suitable groups of features contain variables highly correlated with the feature to be predicted, and are uncorrelated with each other, the CFS approach evaluates the value of subsets of features. Thereby, a matrix of feature-class and feature-feature correlations from the training data is calculated first. The degree of intercorrelation between two features or the correlation between feature X and class Y , which is in the range of $[0,1]$ is quantified as:

$$r_{x,y} = 2.0 \times \left[\frac{\text{gain}}{H(X)+H(Y)} \right] \quad (1)$$

where $\text{gain} = H(X) + H(Y) - H(X,Y)$ represents the information gain between features and classes, and $H(X)$ is the entropy of the feature (Liu et al. 2002). Subsequently, a score value is assigned by using a heuristic in the form of:

$$\text{Merit}_S = \frac{k \overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (2)$$

where Merit_S is the heuristic of a feature subset S with k features, $\overline{r_{cf}}$ represents the average feature-class correlation, and $\overline{r_{ff}}$ represents the average feature-feature inter-correlation. The numerator gives an indication of how predictive a set of features is and the

denominator reveals the redundancy among them. Subsets that contain irrelevant features (with low feature-class correlation $\overline{r_{cf}}$) and features with a high redundancy (high feature-feature correlation $\overline{r_{ff}}$) are evaluated as unsuitable subsets of the feature space (Liu & Schumann 2005). Since features are threatened independently, the CFS approach is not able to identify strongly interacting features. Nonetheless, empirical studies revealed that this method is able to identify useful features under moderate levels of interaction (Hall & Holmes 2003). We used the greedy stepwise search, where the CFS approach starts from the empty set of features with a stopping criterion of five consecutive fully expanded non-improving subsets (preliminary experiments showed that the result of feature selection in this study is not sensitive with respect to the specification of the stopping criterion). Then the subset with the highest merit within an interval of [0, 1] revealed during the search, was selected (Liu et al. 2002).

REGRESSION ANALYSIS

To further explore the strength, direction and significance of the features derived from remote sensing data regarding the dependent variable (*in situ* assessed seismic building vulnerability), we deployed simple ordinary least squares (OLS) regression models for the scoring approach and generalized ordered logit regression models to examine the influence of the features on the respective EMS-98 classes. The basic linear regression model takes the form:

$$y = \alpha + \beta x + \varepsilon \quad (3)$$

with y being the depended variable, x the independent variable, α represents the intercept, β expresses the slope of the relationship between the two variables, and ε is an error term. Regarding the situation where there is more than one independent variable, the regression model is typically expressed as follows:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n + \varepsilon \quad (4)$$

where β_0 is the intercept and $\beta_1 - \beta_n$ represent the slope coefficients for the independent variables $x_1 - x_n$ (see e.g. Kleinbaum et al. 1998 and Montgomery et al. 2001 for a thorough discussion of the simple and multi-linear case). The OLS method minimizes the sum of squared vertical deviations between the observed values in the data set and the values

predicted by the linear approximation. Therefore, the regression line describes, as close as possible, the original values of the dependent variable.

When the dependent variable has a categorical character, the probability of falling into the categories l to i of the dependent variable with I categories is set in relation to the probability of falling in the categories $i+1$ to I :

$$\text{Logit}(Y_{1\dots i/i+1\dots I}|\mathbf{X}) = b_{0i} - b_l X_l - \dots - b_k X_k \quad (5)$$

To take into account the ordinal character of the EMS-98 classification, a generalized ordered logistic regression model is used (see [Williams 2006](#) for a detailed description). The main advantage of generalized ordered logistic regression over ordinal logistic regression is that one does not have to make the assumption that the influence of a predictor is the same for each stage (proportional-odds or parallel-slopes assumption) ([Peterson and Harrell 1990](#)). The applied partial proportional odds model tests the parallel assumption for each of the independent variables and calculates unique beta-coefficients for those violating the assumption. Generalized ordered logistic regression models are therefore less restrictive than ordinal regression models, but more convincing than multinomial logistic regression models. In ordered logistic regressions one single regression constant is calculated for each category of the dependent variable.

The regression analyses are solely applied with the features contained in the subsets with the highest merit as evaluated by the CFS approach. This group of features delivered the best results regarding the supervised regression and classification approaches that are described in the next section. Furthermore, we eliminated collinear variables. To track the composition of the models' performance the explanatory variables were grouped according to the feature categories in [table 2](#) and entered into the regression equations in a stepwise, hierarchical way.

SUPERVISED REGRESSION AND CLASSIFICATION

To predict the respective scoring values and EMS-98 classes, we deploy supervised regression and classification techniques that are based on the delineation of functions from labeled training data. Each instance of the training data is constituted by a dependent variable and an n -dimensional vector of independent variable(s). The supervised learning scheme analyzes the training data and a generalized regression function (for continuous dependent

variables) or classifier (for discrete dependent variables) is delineated to correctly estimate new examples.

SUPERVISED REGRESSION

For estimating the scoring values we compare the merits of multi-linear regression models and Support Vector Machine (SVM) based regression models which are able to represent non-linear boundaries between classes. As described above, multi-linear regression is based on the assumption that the dependent variable Y and its predictors X_1, X_2, \dots, X_n are directly related by a linear combination. Since linear regression models predict poorly in the presence of a nonlinear or non-additive relationship, a nonlinear Support Vector Regression (SVR) approach is additionally utilized. SVMs determine a suitable set of parameters that places a decision surface, the so called hyperplane, between the different classes of training samples according to their position in an n -dimensional feature space. The optimal separating hyperplane is identified as the maximized margin between the different classes and the hyperplane. In a modified form SVMs can also be applied for function estimation (see [Smola and Schölkopf 2004](#)). Detailed theoretical background of SVMs is given in Vapnik ([1995, 1998](#)), Cortes and Vapnik ([1995](#)), and Burges ([1998](#)).

For the calculation of the regression models, the sample data were separated in a stratified manner to use one half of the samples for building the models, and the other half for validation. Regarding the linear regression approach, we excluded collinear features. As a measure of the relative goodness of fit, the Akaike information criterion was used for model selection. For the nonlinear approach we used the Sequential Minimal Optimization algorithm for regression as proposed by Smola and Schölkopf ([1998](#)) with a poly kernel and a regression optimizer introduced by Shevade et al. ([2000](#)). Regarding the comparison of the actual and estimated vulnerability values, we calculated a set of statistical accuracy measures: Mean Error (ME), Mean Percentage Error (MPE), Mean Absolute Error (MAE), Mean Absolute Percentage Error ($MAPE$), Standard Deviation ($StDev$), Relative Standard Deviation ($RStDev$), Pearson product-moment correlation coefficient (R), and $RMSE$.

SUPERVISED CLASSIFICATION

For estimating seismic vulnerability classes according to the EMS-98, we tested several supervised classification techniques (SVMs, radial basis function networks, backpropagation multilayer perceptrons, and random forests) and finally selected a simple naïve Bayes approach, since it outperformed the other classification techniques for a specific set of features. This has been evaluated with respect to the overall accuracy and Cohen's *kappa* statistic (Foody 2002). The simple naïve Bayes approach represents a probabilistic classifier applying Bayes' theorem with strong underlying (naïve) independence assumptions. The basic assumption is the presence (respectively absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable. The aim is to assign an object I to one of a discrete set of categories C_1, C_2, \dots, C_m by using its observable features X_1, X_2, \dots, X_n . The probability of I belonging to a respective category is calculated by applying Bayes' theorem. This is further reduced by assuming mutual conditional independence. Subsequently, I is assigned to the category with the greatest probability (Boyles et al. 2007). For a detailed description see e.g. Duda & Hart (1973) or Lewis (1998).

Analogous to the regression approach, the sample data were separated in a stratified manner, with half of the samples for building the classifier and the other half for its validation. As it can be seen in table 1, the data set used for training the classifier is imbalanced. This is when an uneven distribution of data patterns exists and the number of training instances of a majority class is much larger compared to other minority classes. Hence, the classifier is bias-prone and tends to favor the majority class (Nguyen et al. 2008). To increase the size of the minority classes and balance the class distribution, we oversampled the training data by using the Synthetic Minority Over-sampling Technique (SMOTE; see Chawla et al. 2002). Thereby, new instances are generated based on the "known" distribution to improve the generalization capacity of the learned classifier. The synthetic instances are added in the space between minority examples, emphasizing the class border in favor of the minority class. This principle is applied since the emphasis of class borders is useful in learning efficient discriminative classifiers (Cieslak et al. 2006). As mentioned, the classification outcomes are evaluated by calculating the overall accuracy and the *kappa* statistic. Additionally the user's (precision) and producer's (recall) accuracies as

well as the receiver operating characteristics (ROC Area) for the individual classes are calculated.

RESULTS AND DISCUSSION

FEATURE SELECTION ANALYSIS

SCORING APPROACH

For the scoring approach the Relief-F algorithm reveals 70 features with a relevance value $W(A) > 0$. Due to limited space, only the first 21 most important features are presented what equals the number of features revealed by the CFS approach (Table 3). One can observe similarities regarding the selected features. Most of the features that have a relevance value $W(A) > 0$ are included also in the CFS subset; exceptions include four spectral features of 1st order and two features of 2nd order. Furthermore the features “Roof Type_B”, “Asymmetry_B”, “Perimeter_B”, “Floor Number_B”, “GLCM Homogeneity_{S pan}” and “GLCM Angular 2nd Moment_{S pan}” are merited as very influential by both algorithms, although they have completely different search heuristics.

From the 21 features ranked as most important by the Relief-F approach, 12 are related to buildings, whereas 9 are related to the block level. Thereby, the “Structure Type_s” is merited as most important. This gives an indication that the combined use of physical features is suitable to discriminate homogeneous urban areas that show similar vulnerability characteristics. This feature is followed by roof type and average building height per block. The building height can be considered as an important feature in general, since the floor number of the individual buildings is ranked 7th and represents a feature which is also considered relevant by the CFS approach. The subsequent features that are top-ranked primarily describe the geometry and extent of individual buildings. The merit of the best subset found from the CFS approach is 0.506. The CFS approach reveals a group of features that consists primarily of building level features (16), rather than of block level features (5). Analogous to the Relief-F results, features that characterize the geometry of the individual buildings are included in the set. The CFS subset also includes numerous features that are related to spectral information of 1st and 2nd order. In contrast to the Relief-F results, features that explicitly describe the spatial context were not included.

Table 3. Features revealed for the vulnerability scoring approach using the Relief-F and CFS feature selection

Relief-F (Ranker)			CfsSubsetEval (Greedy Stepwise)	
<i>Value</i>	<i>Feature Name</i>	<i>Feature category</i>	<i>Feature Name</i>	<i>Feature category</i>
0.0435	Structure Type _s	IKONOS - Spatial context	Perimeter _B	IKONOS - Extent
0.0239	Roof Type _B	IKONOS - Shape	Asymmetry _B	IKONOS - Shape
0.0142	Avg. Building Height _s	DSM - 3D Features	Density _B	
0.0110	Width _B	IKONOS - Extent	Roof Type _B	IKONOS - Spectral Information (1 st order)
0.0109	Perimeter _B		M(1)/M(4) _B	
0.0106	Dist. 3 rd Neighbor Building _B	IKONOS - Spatial context	M(2)/M(3) _B	
0.0105	Floor Number _B	DSM - 3D Features	M(2)/M(4) _B	
0.0098	Length _B	IKONOS - Extent	M(1)/M(2) _B	
0.0076	Area _B		M(3)/M(4) _B	
0.0071	Dist. 2 nd Neighbor Building _B	IKONOS - Spatial context	$[M(4)-M(3)] / [M(4) + M(3)]$ _B	
0.0069	GLCM Homogeneity _{s pan}	IKONOS - Spectral Information (2 nd order)	$M(2) / [M(1)+M(2)+M(3)+M(4)]$ _B	
0.0064	Degree of Building Density _s	IKONOS - Spatial context	M(2)/M(4) _s	
0.0062	Share of LC class "sealed" _s		$M(2) / [M(1)+M(2)+M(3)+M(4)]$ _s	
0.0060	Dist. 1 st Neighbor Building _B		$[M(4)/M(3)] / [M(1)+M(2)+M(3)+M(4)]$ _s	
0.0059	Average Building Size _s		GLDV Angular 2 nd Moment _{B pan}	
0.0056	Effective Area _B	DSM - 3D Features	GLDV Angular 2 nd Moment _{B nir}	IKONOS - Spectral Information (2 nd order)
0.0055	Area Building Block _s	IKONOS - Spatial context	GLCM Angular 2 nd Moment _{s pan}	
0.0052	$[M(3)-M(1)] / [M(3)+M(1)]$ _s	IKONOS - Spectral Information (2 nd order)	GLCM Homogeneity _{B pan}	
0.0045	Asymmetry _B	IKONOS - Shape	GLCM Homogeneity _{B nir}	
0.0042	GLCM Angular 2 nd Moment _{s pan}	IKONOS - Spectral Information (2 nd order)	GLCM Homogeneity _{s pan}	
0.0042	GLCM Standard Deviation _{B pan}		Floor Number _B	DSM - 3D Features

EMS-98 CLASSIFICATION

Regarding the EMS-98 data set, the Relief-F approach scored only eight features with a value $W(A) < 0$. The best CFS subset has a merit of 0.106 and consists of nine features (Table 4), whereby these features all have a positive $W(A)$ value. The features “Structure Type s ”, “Floor Number B ”, “Std. Dev. Blue(1) s ”, “Compactness B ”, “Share of Impervious Surface s ”, “Std. Dev. Red(3) s ”, which are part of CFS subset are also ranked as very influential by the Relief-F approach. The majority of features refer to the block level for both methods. Similar to the results of the scoring approach, the “Structure Type s ” is merited as most important by the Relief-F approach, followed by the estimated period of construction. Again, the height characteristics of the urban environment play an important role, since the individual height of buildings and the average building height per block are considered very influential. The CFS subset contains features related to geometry and spectral characteristics. Additionally, features which describe the spatial context where the buildings are embedded in, are given notable scores.

Table 4. Features revealed for the EMS-98 classification using the Relief-F and CFS approach.

Relief-F (Ranker)			CfsSubsetEval (Greedy Stepwise)	
<i>Value</i>	<i>Feature Name</i>	<i>Feature category</i>	<i>Feature Name</i>	<i>Feature category</i>
0.0589	Structure Type s	IKONOS - Spatial context	Area B	IKONOS – Extent
0.0212	Period of construction s	LANDSAT – TM, ETM+	Compactness B	IKONOS – Shape
0.015	Roof Type B	IKONOS - Shape	Std. Dev. Blue(1) s	IKONOS - Spectral Information (1 st order)
0.0143	Avg. Building Height s	DSM - 3D Features	Std. Dev. Red(3) s	
0.0142	Main Direction B	IKONOS - Shape	GLCM Entropy B_{pan}	IKONOS - Spectral Information (2 nd order)
0.0132	GLCM Std. Dev. s_{pan}	IKONOS - Spectral Information (2 nd order)	Average Building Size s	IKONOS - Spatial context
0.0124	Floor Number B	DSM - 3D Features	Share of Impervious Surface s	
0.0113	$[M(3)-M(1)]/[M(3)+M(1)]_s$	IKONOS - Spectral Information (1 st order)	Structure Type s	
0.0112	GLCM Homogeneity s_{pan}	IKONOS - Spectral Information (2 nd order)	Floor Number B	DSM - 3D Features

REGRESSION ANALYSIS

SCORING APPROACH

Results from the OLS models (table 5) indicate that the building height and features that describe the geometry of individual buildings are the most important remotely sensed determinants of seismic building vulnerability, with significance influence at the 99.9 percent level. The combined use of these features results in an adjusted coefficient of determination of 0.254, which covers approximately 85% of the explained share of variance. The additional features related to 1st and 2nd order spectral information alone, can raise the model's R^2 value to 0.299. This finding is reasonable since the first mentioned features are reflected by comparable features part of the scoring value calculation itself, although for example, the geometry parameters are approximated mathematically in a completely different manner.

EMS-98 CLASSIFICATION

Table 6 summarizes the results of the generalized ordered logistic model. It shows that having a higher value of "Area_B" increases the likelihood of being in class B-C or C-D-E, and therefore to be less vulnerable. Compact structures are more likely to be assigned to a class of higher vulnerability. In contrast, Low Income Structure Types (Structure Type_S = Low Income) again have a higher chance of belonging to a class of lower vulnerability in comparison to slum structure types (note that no differences are observed between other structure types and the slum structure). Unique beta-coefficients are calculated for the influence of the "Floor Number_B" on the different EMS-98 classes. This indicates that the parallel assumption was not met for this variable. Having a higher "Floor Number_B" increases the chance of being in EMS-98 group A-B in comparison to the others. An even stronger negative effect of "Floor number_B" is found for the probability of belonging to EMS-98 class B-C.

The latter finding is contrary to the effect of "Floor Number_B" regarding the scoring approach. However, this observation reflects oppositional assumptions associated with the different assessment approaches. Regarding the scoring approach, an increase in building height increases the value, implying lower vulnerability. Based on this it was assumed that high rise buildings represent engineered structures that are built in a more sophisticated way than informal (low rise) settlements. For the EMS-98 there were no such unidirectional assumptions made and the analysis results indicate that higher buildings are more likely to be

more vulnerable. This is consistent with observations derived from damage data collected after the $M_w = 7.6$ event on September 30th, 2009 ([Sengara et al. 2010](#)).

Table 5. Results of simple OLS models regarding the scoring approach (coefficients and standard error). The grouped features were entered into the regression equations in a stepwise hierarchical way.

Feature	Model 1		Model 2		Model 3		Model 4	
	Coef.	Std. Err.						
Floor Number _B	1.034508***	.1132338	.7293502***	.1144819	.6392977***	.11527	.5193538***	.1165116
Perimeter _B			.0090334***	.0024904	.0087243***	.0024959	.0085121***	.0025175
Asymmetry _B			3.31487***	.888291	3.354441***	.8640369	3.112394***	.8651691
Density _B			5.073578***	1.123404	5.063067***	1.081441	4.660547***	1.063173
Roof Type _B			.1545023	.2558068	.1839185	.2544025	.2649701	.2615001
GLCM Homogeneity _{B pan}					1.07571	1.315121	.4871516	1.336938
GLCM Homogeneity _{B nir}					1.350518*	.5738469	.7967434	.6515387
GLCM Angular 2 nd Moment _{S pan}					396.783*	170.6297	190.7972	217.2341
GLCM Homogeneity _{S pan}					-11.61803	9.150705	-1.461658	10.16228
M(3)/M(4) _B							1.167996	1.689854
M(2)/[M(1)+M(2)+M(3)+M(4)] _B							12.98874*	6.32278
[M(4)/M(3)]/ [M(1)+M(2)+M(3)+M(4)] _S							-29.81346	105.0541
M(2)/M(4) _S							1.475905	1.412028
Number of observations		434		434		434		434
adj. R ²		0.165		0.254		0.273		0.299

*P < 5%, **P < 1%, ***P < 0.1%

Table 6. Results of generalized ordered logit models regarding the EMS-98 classification (coefficients and standard error). First section refers to the relation of class A-B and C-D-E; second section refers to the relation of class B-C and C-D-E (only relevant for “Floor Number_B”). The grouped features were entered into the regression equations in a stepwise hierarchical way.

Feature	Model 1		Model 2		Model 3		Model 4	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Floor Number _B	.4678125***	.0929906	.3403922***	.0980104	.2258272*	.1060161	.1886913+	.1086677
Area _B			.0006703***	.0001668	.0006324***	.0001679	.0005222*	.000228
Compactness _B			-1.604815***	.4134367	-1.44446***	.415815	-1.51715***	.4242106
Std. Dev. Blue(1) _S					.0344184**	.0128353	.0272497+	.0154555
GLCM Entropy _{B pan}					.0190522	.0723382	.0119984	.0725885
Average Building Size _S							.0010974	.0009604
Share of Impervious Surface _S							.0203762	.0211656
Structure Type _S = Suburb							.7969139	.5182687
Structure Type _S = Low Income							.9920697*	.4511381
Structure Type _S = Medium Income							.6634897	.4319753
Structure Type _S = High Income							.8390549	.703233
Floor Number _B	-.0793007	.1457927	-.2491776	.155034	-.3606772*	.1624929	-.3851584*	.1644611
Area _B			.0006703***	.0001668	.0006324***	.0001679	.0005222*	.000228
Compactness _B			-1.604815***	.4134367	-1.44446***	.415815	-1.51715***	.4242106
Std. Dev. Blue(1) _S					.0344184**	.0128353	.0272497+	.0154555
GLCM Entropy _{B pan}					.0190522	.0723382	.0119984	.0725885
Average Building Size _S							.0010974	.0009604
Share of Impervious Surface _S							.0203762	.0211656
Structure Type _S = Suburb							.7969139	.5182687
Structure Type _S = Low Income							.9920697*	.4511381
Structure Type _S = Medium Income							.6634897	.4319753
Structure Type _S = High Income							.8390549	.703233
<i>Number of observations</i>		434		434		434		434

*P < 5%, **P < 1%, ***P < 0.1%

SUPERVISED REGRESSION AND CLASSIFICATION

SUPERVISED REGRESSION

The multi-linear and SVR models are calculated based on (i) the originate 132-dimensional feature vector, (ii) the 70 features with a positive degree of relevance as evaluated by the Relief-F algorithm, (iii) the group of 21 features with the highest merit as revealed by the CFS approach, and (iv) the further reduced set of features as used for the regression analysis (RA) (collinear features were removed from the subsets (i),(ii), and (iii) for the linear models). The evaluated model estimation results are given in table 7.

Table 7. Evaluated results of the model predictions of different feature sets compared to reference values.

Multi-linear Regression								
Used features	<i>MAE</i>	<i>MAPE</i>	<i>ME</i>	<i>MPE</i>	<i>StDev</i>	<i>RStDev</i>	<i>R</i>	<i>RMSE</i>
All features	2.33	13.32	-0.56	-3.02	2.56	14.32	0.43	2.98
Relief-F $W(A) > 0$	2.02	11.84	-0.11	-0.58	2.07	11.29	0.48	2.62
CFS subset	1.84	10.99	-0.17	-0.91	1.52	8.33	0.56	2.41
CFS subset (RA)	1.80	10.81	-0.20	-1.10	1.45	7.96	0.58	2.37
Support Vector Regression								
Used features	<i>MAE</i>	<i>MAPE</i>	<i>ME</i>	<i>MPE</i>	<i>StDev</i>	<i>RStDev</i>	<i>R</i>	<i>RMSE</i>
All features	1.88	11.26	0.02	0.13	1.91	10.37	0.53	2.48
Relief-F $W(A) > 0$	1.79	11.07	0.43	2.31	1.72	9.13	0.57	2.42
CFS subset	1.72	10.61	0.23	1.25	1.54	8.28	0.59	2.36
CFS subset (RA)	1.73	10.67	0.24	1.31	1.63	8.73	0.59	2.36

When using the CFS feature set and the SVR approach the best results are achieved, with lowest *MAE* (1.72) / *MAPE* (10.61) and *RMSE* (2.36) and highest linear correlation ($R = 0.59$) of the model estimates. This demonstrates the viability of the approach. Furthermore, the nonlinear approach shows better predictions compared to the linear approach regarding the respective set of features used. The actual vulnerability values and the predicted values for the best model are visualized in a scatter plot (Fig. 4). One can observe that the model overestimates low scoring values, buildings with a high seismic vulnerability - and underestimates high scoring values - characterizing buildings with a low seismic vulnerability. This observation is also true for all other models calculated.

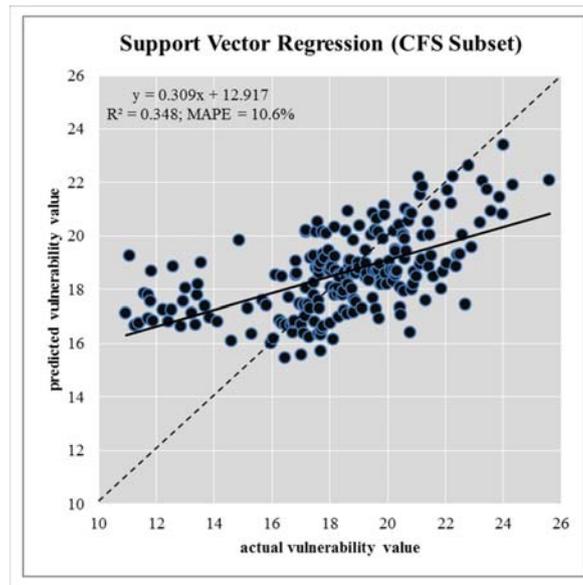


Figure 4. Scatter plot of the actual and predicted scoring values (SV) by using the CFS feature subset and the SVR approach.

SUPERVISED CLASSIFICATION

As mentioned above, the best result of all classification approaches in combination with different feature sets evaluated (analogous to the different feature sets used for building the supervised regression models), was achieved with the naïve Bayes classifier and the features revealed by the CFS approach. With this set of features only, the classifier performed considerably better than the other classification approaches evaluated. This is most likely attributable to the low pair-wise correlation of the features in this subset.

The overall accuracy for the classification is 65.4% with a “fair” (Landis & Koch 1977) *Kappa* statistic of 0.36. The classification performance for the respective EMS-98 classes is shown in table 8. One can find the classifications for the most vulnerable buildings according to the EMS-98 classification (class A-B) to be the most accurate with a classification accuracy of 0.696 (user’s) and 0.793 (producer’s), respectively. In contrast, the classification outcomes for the least vulnerable structures according to the EMS-98 classification (class C-D-E) are less feasible, with a user’s accuracy of 0.294 and a producer’s accuracy of 0.227, and the lowest ROC Area values. The strong confusion and poor predictive performance regarding class C-D-E is most likely caused by the small number of instances and high intra-class variability.

Table 8. Results for the EMS-98 classes of the supervised classification approach using the CFS Subset features.

Class	User's Accuracy (Precision)	Producer's Accuracy (Recall)	ROC Area
A-B	0.696	0.793	0.666
B-C	0.661	0.554	0.686
C-D-E	0.294	0.227	0.601
Weighted Average	0.643	0.654	0.645

CONCLUSIONS

This study shows that an indirect correlation between physical information in the (urban) environment, drawn from remote sensing data and seismic vulnerability of buildings, exists. We demonstrated how to derive and identify meaningful features using a combination of remote sensing data and how to quantitatively evaluate their explanatory power. By means of a sequential procedure of calculating features from very high resolution multispectral data, height information and spatiotemporal analyses, and of applying machine learning approaches, the seismic vulnerability of buildings can be estimated with viable accuracies. Furthermore, we identified features that have high explanatory content and are most useful for the estimation of seismic building vulnerability, in terms of a preliminary screening. The features merited as most useful, their influence and direction vary considerably in dependency of the *in situ* seismic vulnerability assessment applied. However, features such as building height characteristics and features related to the geometry of the individual buildings turned out to generally explain seismic vulnerability to a significant degree.

We believe that this study generated insight into the capabilities of remote sensing for assessing seismic building vulnerability, and can be helpful for further empirical case studies in other parts of the world. Future research activities should comprise the systematic evaluation of remote sensing data according to wide-spread vulnerability assessment methods, such as the EMS-98, to make results comparable. In addition, the evaluation of the capabilities of remote sensing for estimating the seismic building structural type is a crucial task (see e.g. [Sarabandi & Kiremidjian 2007](#) for related work). Latest and future spaceborne missions, such as WorldView-2 and 3, CARTOSAT-3 or ALOS-3 have enhanced resolution characteristics, which will allow for the calculation of more sophisticated feature vectors ([Novack et al. 2011](#)). It is believed that this will entail more accurate predictive performance

of the models and further increase the applicability of remote sensing to assess the seismic vulnerability of buildings.

We conclude that remote sensing data and methods have a high capability to support large area assessments of building vulnerability, indicating the need for systematic application and validation of our findings. Lastly, we want to trigger an open dialogue between the remote sensing and earthquake engineering community, to better understand how remote sensing data can be linked to assessment approaches from engineering science in a robust, standardized and transferable way, to define common scales and enable systematic large-area assessments and monitoring of dynamic earthquake prone urban areas around the globe.

ACKNOWLEDGMENTS

The research was founded partially by Helmholtz-EOS (Earth Observation System). The authors wish to thank the three anonymous reviewers for their very constructive comments and suggestions and Sarah Banks for her help with proofreading the manuscript.

REFERENCES

- Bilham, R., 2009. The seismic future of cities, *Bulletin of earthquake engineering* **7**, 839–887.
- Borfecchia, F., Pollino, M., De Cecco, L., Lugari, A., Martini, S., La Porta, L., Ristatore, E., and Pascale, C., 2010. Active and passive remote sensing for supporting the evaluation of the urban seismic vulnerability, *Italian Journal of Remote Sensing* **42**, 129–141.
- Borzi, B., Dell'Acqua, F., Faravelli, M., Gamba, P., Lisini, G., Onida, M., and Polli, D., 2011. Vulnerability study on a large industrial area using satellite remotely sensed images, *Bulletin of Earthquake Engineering* **9**, 675–690.
- Boyles, S., Fajardo, D., and Waller, S. T. (2007). A Naïve Bayesian Classifier for Incident Duration Prediction. In *Proceedings of the 86th Annual Meeting of the Transportation Research Board*. Washington D.C., USA
- Burges, C.J.C., 1998. A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge discovery* **2**, 1–47.
- Calvi, G.M., Pinho, R., Magenes, G., Bommer, J.J., Restrepo-Vélez, L.F., and Crowley, H., 2006. Development of seismic vulnerability assessment methodologies over the past 30 years, *ISET Journal of Earthquake Technology* **43**, 75–104.
- Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* **16**, 321–357
- Cieslak, D. A., Chawla, N. V., and Striegel, A. (2006). Combating imbalance in network intrusion datasets. In *Proceedings of the IEEE International Conference on Granular Computing*, pp. 732–737.
- Cortes, C., and Vapnik, V., 1995. Support Vector Networks, *Machine Learning* **20**, 1–25.

- Crowley, H., Pinho, R., and Bommer, J.J., 2004. A probabilistic displacement-based vulnerability assessment procedure for earthquake loss estimation, *Bulletin of Earthquake Engineering* **2**, 173–219.
- Deichmann, U., Ehrlich, D., Small, C., and Zeug, G., 2011. Using high resolution satellite data for the identification of urban natural disaster risk. JRC, GFDRR, IPSC, WB, 80pp.
- Duda, R., Hart, P., 1973. Pattern Classification and Scene Analysis, John Wiley and Sons, New York.
- Freeman, A.S., 2004. Review of the development of the capacity spectrum method, *ISSET Journal of Earthquake Technology* **41**, 1–13.
- Foody, G.M., 2002. Status of land cover classification accuracy assessment. *Remote Sensing of Environment* **80**, 185-201
- French, S.P., and Muthukumar, S., 2006. Advanced technologies for earthquake risk inventories, *Journal of Earthquake Engineering* **10**, 207–236.
- Geiß, C., Taubenböck, H., 2013. Remote sensing contributing to assess earthquake risk: from a literature review towards a roadmap. *Natural Hazards*. Doi: 10.1007/s11069-012-0322-2
- GEM, 2013. Global Earthquake Model, Inventory Data Capture Tools. URL: <http://www.globalquakemodel.org/risk-global-components/inventory-capture> Accessed 14 Feb 2013
- Grünthal, G., Musson, R., Schwarz, J., and Stucchi, M., 1998. European Macroseismic Scale, *Cahiers de Centre Européen de Géodynamique et de Seismologie*, vol 15, Luxembourg.
- Hall, M., and Holmes, G., 2003. Benchmarking attribute selection techniques for discrete class data mining, *IEEE Transactions on Knowledge and Data Engineering* **15**, 1–16.
- Haralick, R., Shanmugam, K., and Dinstein, I., 1973. Textural features for image classification, *IEEE Transactions on Systems, Man, and Cybernetics* **SMC-3**, 610–621.
- Herold, M., Liu, X., and Clarke, K., 2003. Spatial metrics and image texture for mapping urban land use, *Photogrammetric Engineering and Remote Sensing* **69**, 991-1001.
- Hughes, G.F., 1968. On the mean accuracy of statistical pattern recognizers. *IEEE Transactions on Information Theory* **14**(1): 55–63
- Irish, R., 2008. Landsat 7 Science data user's handbook, *Landsat Project Science Office*, Issue 21, 186 pp.
- Jaiswal, K., Wald, D., D'Ayala, D., 2011. Developing Empirical Collapse Fragility Functions for Global Building Types. *Earthquake Spectra* **27**(3):775-795
- Kleinbaum, D.G., Kupper, L.L., Mueller, K.E., Nizam, A., 1998. Applied regression analysis and other multivariable methods, 3rd ed., Duxbury Press, Pacific Grove, CA, 736pp.
- Kononenko, I., 1994. Estimating attributes: Analysis and extension of RELIEF, in *Proceedings of the European Conference on Machine Learning*, Springer, Catania, Italy, pp. 171–182
- Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. *Biometrics* **33**, 159-174.
- Lewis, D.D., 1998. Naïve Bayes at forty: The independence assumption in information retrieval, *Proceedings of ECML*.
- Li, X., Tennant, K., and Lawrence, G., 2004. Three-dimensional mapping with airborne IFSAR based STAR technology – Intermap's experiences, *International Archives Photogrammetry and Remote Sensing* Vol. XXXV-B3, pp. 261–266
- Liu, H., Li, J., and Wong, L., 2002. A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns, *Genome Informatics* **13**, 51–60.
- Liu, Y., and Schumann, M., 2005. Data mining feature selection for credit scoring models, *Journal of the Operational Research Society* 1-10.

- Maktav., D., Erbek, F.S., and Jürgens, C., 2005. Remote sensing of urban areas, *International Journal of Remote Sensing* **26**, 655–659.
- McCloskey, J., Lange, D., Tilmann, F., Nalbant, S.S., Bell, A.F., Natawidjaja, D.H., and Rietbrock, A., 2010. The September 2009 Padang earthquake, *Nature Geoscience* **3**, 70–71.
- Montgomery, D.C., Peck, E.A., and Vining, G.G., 2001. *Introduction to Linear Regression Analysis*, 3rd edition, Wiley, New York, NY, 672 pp.
- Mück, M., Taubenböck, H., Post, J., Wegscheider, S., Strunz, G., Sumaryono, S., Ismail F.A., 2013. Assessing building vulnerability to earthquake and tsunami hazard using remotely sensed data. Accepted for publication in *Natural Hazards*
- Mueller, M., Segl, K., Heiden, U., and Kaufmann, H., 2006. Potential of high-resolution satellite data in the context of vulnerability of buildings, *Natural Hazards* **38**, 247–258.
- Nguyen, G.H., Bouzerdoum, A., Phung, A., 2008. A supervised learning approach for imbalanced data sets, *Pattern Recognition*, pp. 1–4
- Novack, T., Esch, T., Kux, H., and Stilla, U., 2011. Machine learning comparison between WorldView-2 and QuickBird-2-Simulated imagery regarding object-based urban land cover classification, *Remote Sensing* **3**, 2263–2282.
- Pacifici, F., Chini, M., Emery, W., 2009. A neural network approach using multi-scale textural metrics from very high resolution panchromatic imagery for urban land-use classification, *Remote Sens. Environ.*, **113**(6), 1276–1292.
- Peterson, B.L., Harrell, F.E.J., 1990. Partial proportional odds models for ordinal response variables, *Applied Statistics*, **39**(2), 205–217.
- Polli, D., and Dell'Acqua, F., 2011. Fusion of optical and SAR data for seismic vulnerability mapping of buildings, in *Optical remote sensing. Advances in signal processing and exploitation techniques*, Prasad, S., Bruce, L.M., and Chanussot J, Eds, Springer, Heidelberg, 329–341
- Prasad, J.S.R., Singh, Y., Kaynia, A.M., and Lindholm, C., 2009. Socioeconomic clustering in seismic risk assessment of urban housing stock, *Earthquake Spectra* **25**, 619–641.
- Rashed, T., and Jürgens, C., 2010. *Remote sensing of urban and suburban areas. Remote Sensing and Digital Image Processing*, Vol. 10, 1st Edition, XII, 352 pp.
- Richter, R., 1996. A spatially adaptive fast atmospheric correction algorithm, *International Journal of Remote Sensing* **17**, 1201–1214.
- Robnik-Šikonja, M., and Kononenko, I., 2003. Theoretical and empirical Analysis of ReliefF and RReliefF, *Machine Learning Journal* **53**, 23–69.
- Sahar, L., Muthukumar, S., and French, P., 2010. Using aerial imagery and GIS in automated building footprint extraction and shape recognition for earthquake risk assessment of urban inventories, *IEEE Transactions on Geoscience and Remote Sensing* **48**, 3511–3520.
- Sarabandi, P., Kiremidjian, A., 2007. Development of algorithms or building inventory compilation through remote sensing and statistical inferencing. The John A. Blume Earthquake Engineering Center, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA, Report No. 158, 419p.
- Sarabandi, P., Kiremidjian, A., Eguchi, R. T. and Adams, B.J., 2008. Building inventory compilation for disaster management: Application of remote sensing and statistical modeling. *Technical Report Series MCEER-08-0025*, MCEER: Buffalo.
- Sengara, I.W., Suarjana, M., Beetham, D., Corby, N., Edwards, M., Griffith, M., Wehner, M., Weller, R., 2010. The 30th September 2009 West Sumatra Earthquake. Padang Region Damage Survey. Geoscience Australia, Record 2010/44, 201pp.

- Shevade, S.K., Keerthi, S.S., Bhattacharyya, C., and Murthy, K.R.K., 2000. Improvements to the SMO algorithm for SVM regression, *IEEE Transactions on Neural Networks* **11**, 1188–1193.
- Sirmacek, B., Taubenböck, H., Reinartz, P., Ehlers, M., 2012. Performance evaluation for 3-D city model generation of six different DSMs from air- and spaceborne sensors, *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **5**, 59–70.
- Smola, A.J., and Schölkopf, B., 1998. A tutorial on support vector regression, *NeuroCOLT2 Technical Report Series - NC2-TR-1998-030*
- Smola, A.J., and Schölkopf, B., 2004. A tutorial on support vector regression, *Statistics and Computing* **14**, 199–222.
- Taubenböck, H., Post, J., Kiefl, R., Roth, A., Ismail, F.A., Strunz, G., and Dech, S., 2008a. Risk and vulnerability assessment to tsunami hazard using very high resolution satellite data, in *Proceedings of the EARSeL Joint Workshop: Remote Sensing: New Challenges of high resolution*, Carsten, J. (editor), Bochum, Germany
- Taubenböck, H., Post, J., Roth, A., Strunz, G., Kiefl, R., Dech, S., and Ismail, F., 2008b. Multi-scale assessment of population distribution utilizing remotely sensed data – The case study Padang, West Sumatra, Indonesia, *International Conference on Tsunami Warning (ICTW)*, Bali, Indonesia, 12-14 November 2008
- Taubenböck, H., Roth, A., Dech, S., Mehl, H., München, J.C., Stempniewski, L., and Zschau, J., 2009a. Assessing building vulnerability using synergistically remote sensing and civil engineering, in *Urban and Regional Data Management*, Kreck, A., Rumor, M., Zlatanova, S., Fendel, E., Eds, Taylor & Francis Group, London, 287–300
- Taubenböck, H., Goseberg, N., Setiadi, N., Lämmel, G., Moder, F., Oczipka, M., Klüpfel, H., Wahl, R., Schlurmann, T., Strunz, G., Birkmann, J., Nagel, K., Siegert, F., Lehmann, F., Dech, S., Gress, A., and Klein, R., 2009b. « Last-Mile » preparation for a potential disaster – Interdisciplinary approach towards tsunami early warning and an evacuation information system for the coastal city of Padang, Indonesia. *Natural Hazards and Earth System Science* **9**, 1509–1528.
- Taubenböck, H., Wurm, M., Setiadi, N., Gebert, N., Roth, A., Strunz, G., Birkmann, J., and Dech, S., 2009c. Integrating remote sensing and social science – The correlation of urban morphology with socioeconomic parameters, *Urban Remote Sensing Joint Event*, Shanghai, China
- Taubenböck, H., Goseberg, N., Lämmel, G., Setiadi, N., Schlurmann, T., Nagel, K., Siegert, F., Birkmann, J., Traub, K.-P., Dech, S., Keuck, V., Lehmann, F., Strunz, G., Klüpfel, H., 2013. Risk reduction at the « Last-Mile » : an attempt to turn science into action by the example of Padang, Indonesia, *Natural Hazards* **65**(1), 915-945
- Vapnik, V.N., 1995. *The nature of statistical learning theory*, Springer, New York, 187 pp.
- Vapnik, V.N., 1998. *Statistical learning theory*, *Wiley Series on Adaptive and Learning Systems*, Wiley, New York
- Weszka, J.S., Dyer, C.R., and Rosenfeld, A., 1976. A comparative study of texture measures for terrain classification, *IEEE Transactions on Systems, Manufacturing and Cybernetics* **SMC-6**, 2269–2285.
- Weng, Q., and Quattrochi, D.A., 2006. *Urban remote sensing*, CRC Press/Taylor and Francis, Boca Raton, FL, USA, 432 pp.
- Whitman, R.V., 1973. *Damage probability matrices for prototype buildings*, Structures Publication, Department of Civil Engineering, Massachusetts Institute of Technology, Boston, Massachusetts, USA
- Wieland, M., Pittore, M., Parolai, S., Zschau, J., Moldobekov, B., and Begaliev, U., 2012. Estimating building inventory for rapid seismic vulnerability assessment: Towards an integrated

approach based on multi-source imaging, *Soil Dynamics and Earthquake Engineering* **36**, 70-83.

Williams, R., 2006. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *The Stata Journal* **6**(1), 58-82.

Wu, S.-S., Qiu, X., Usery, E.L., and Wang, L., 2008. Using geometrical, textural, and contextual information of land parcels for classification of detailed urban land use. *Annals of the Association of American Geographers* **99**, 76–98.