



# The necessary yet complex evaluation of 3D city models: a semantic approach

Oussama Ennafii, Clément Mallet, Arnaud Le Bris, Florent Lafarge

## ► To cite this version:

Oussama Ennafii, Clément Mallet, Arnaud Le Bris, Florent Lafarge. The necessary yet complex evaluation of 3D city models: a semantic approach. JURSE 2019 – Joint Urban Remote Sensing Event, May 2019, Vannes, France. hal-02165562v1

**HAL Id: hal-02165562**

**<https://inria.hal.science/hal-02165562v1>**

Submitted on 26 Jun 2019 (v1), last revised 17 Jul 2019 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# The necessary yet complex evaluation of 3D city models: a semantic approach

Oussama Ennafii, Clément Mallet, Arnaud Le Bris  
Univ. Paris Est, IGN-ENSG, LaSTIG, Saint-Mandé, France  
firstname.lastname@ign.fr

Florent Lafarge  
Inria, TITANE, Sophia Antipolis, France  
florent.lafarge@inria.fr

**Abstract**—The automatic modeling of urban scenes in 3D from geospatial data has been studied for more than thirty years. However, the output models still have to undergo a tedious task of correction at city scale. In this work, we propose an approach for automatically evaluating the quality of 3D building models. A taxonomy of potential errors is first proposed. Handcrafted features are computed, based on the geometric properties of buildings and, when available, Very High Resolution images and depth data. They are fed into a Random Forest classifier for the prediction of the quality of the models. We tested our framework on three distinct urban areas in France. We can satisfactorily detect, on average 96% of the most frequent errors.

**Index Terms**—3D, modeling, buildings, quality, taxonomy, classification, error detection, geometry, VHR data.

## I. INTRODUCTION

Although automatic 3D urban modeling has been extensively studied for many years by academics and industrials alike, current algorithms fail to adapt to different urban settings [1]. Output building models still have to be inspected by human operators for quality assessment and subsequent correction. Thus, scalability is not handled in the literature albeit being critical in operational contexts. In this work, we study the automation of 3D building model evaluation. We focus on assessing polyhedral structured models. They represent building roof architectures, which result from a given urban reconstruction method, unknown here. Each facet of the model corresponds to an architectural (post of the time planar) feature. Such 3D models do not correspond to triangle meshes. The latter ones exhibit higher data fidelity but lower compacity and no semantics. Depending on the spatial resolution of input data, the urban scene, and the intended use, the reconstituted model has a certain **Level of Detail (LoD)** [2].

Few works have addressed the quality assessment of 3D urban models at building level. Usually, visual inspection [3] or comparison with field measurements [4] are performed, void of structural information. Our work focuses on semantically evaluating the structural compliance of modeled buildings. We define an evaluation framework that can be used for building model correction, change detection, reconstruction method selection, and crowdsourcing evaluation. We ignore format/geometric issues [5]. Our framework targets to be independent from the Level of Detail and the modeling method. Quality metrics can be provided with the reconstruction process but lack of independence with the underlying method.

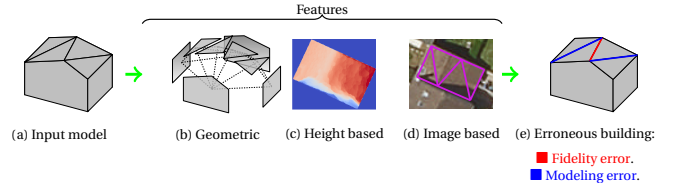


Fig. 1: The semantic evaluation paradigm proposal. Geometric structural features are first proposed (b), potentially augmented with attributes based on height map and color gradient comparison (c-d). Semantic errors are eventually predicted using a pretrained classifier.

This work proposes an adaptable and flexible framework. It is indifferent to input urban scenes and reconstruction methods (Figure 1). For that purpose, our contributions are three-fold:

- A new **taxonomy of errors**, hierarchical, adapted to all LoD, and independent from input models;
- A **supervised classification** formulation of the problem, which predicts all errors affecting the building model;
- A multimodal **baseline of features**, which are extracted both from the model and external data (Very High Resolution optical images and height data).

Section II introduces the problem of the evaluation of 3D building models and discusses existing methods. Section III details the proposed approach, while data and results of experiments conducted over three urban areas are presented in Section IV. Main conclusions are drawn in Section V.

## II. RELATED WORK

3D urban models qualification has been barely investigated in the literature despite plethora of modeling approaches. Methods can be classified based on the assessment criteria: geometric fidelity indices or semantic errors. Geometric fidelity metrics quantify the modeling accuracy relying on the precision of points of interest (e.g., intersection points), of surfaces or on the 3D model volume. These are usually compared to higher resolution reference data [4], [6]. These metrics fail to describe local modeling defects. Consequently, semantic errors are preferred for such a task, often following the traffic light paradigm. For instance, four levels of quality are defined in [7]: Correct, Acceptable, Generalized, and False. However, these levels are not explicitly quantified and are ill

defined: they depends on the end-user. The error taxonomy can also adopt the modeling method perspective. Errors can be grouped into footprint ones (e.g., erroneous outline, missing inner court) and modeling errors (under/over-segmentation, height inaccuracy). These are therefore method-dependent. The building model is therefore assessed owing to a supervised classification that takes these defined errors as labels. In order to represent input models, features are extracted from aerial images or Digital Surface Model (DSM) at very high resolution (20 cm to 25 cm), by 3D segments comparisons or texture correlation ratios [7], [8]. Despite satisfactory results, this approach exhibits two main drawbacks: a taxonomy specific to some urban scenes or modeling methods and a complex feature computation step, preventing upscaling strategies. The main idea of our work is to propose an agnostic assessment approach that do not rely on reference data and complex features.

### III. METHODOLOGY

Our method relies on predicting errors in order to qualify a 3D model at the scale of a building. Errors are defined in a hierarchical taxonomy, based on the assessment objectives. Handcrafted features of various kinds are computed (Figure 1). A supervised classifier is trained using annotated building models and tested later on unseen buildings for quality prediction based on the detected errors. This pipeline is modular: it takes into account geometric features, coming from the 3D model, which can be augmented by height features, from the DSM, or optical features, from a VHR orthoimage.

#### A. Error taxonomy

Two criteria are taken into account so as to build a generic and flexible taxonomy (Figure 2): the LoD and the semantic level, which we call henceforth *finesse*. An *atomic* error is one with maximal *finesse*: it corresponds, intuitively, to a unitary action that an operator should take in order to correct a model.

From an operational perspective, not all buildings are qualifiable. For instance, buildings can be occluded by vegetation or happen to show self-intersection problems. These pathological cases are outside the realm of our assessment framework. Thus, in the *finesse* level equal to 0, we discriminate between *Qualifiable* and *Unqualifiable* models. At the next level, models are classified being *Valid* or *Erroneous*. Based on the LoD, these last models are distinguished into error families at *finesse* equal to 2. *Building errors* correspond to defects at the building level ( $\text{LoD-1} \cup \text{LoD-0}$ ). Errors related to building facets ( $\text{LoD-2}$ ) are assembled into *Facet errors*. The last family, *Superstructure errors*, groups errors pertaining to LoD-3 errors. These families contain *atomic* errors that have a maximal *finesse* equal to 3.

This error arrangement does not depend on a particular modeling method or urban environment. Labels are not redundant as *atomic* errors are chosen to be independent from each other. They represent a particular topological or geometric defect. Topological errors relate to structural defects in the model, while geometrical ones address modeling imprecision.

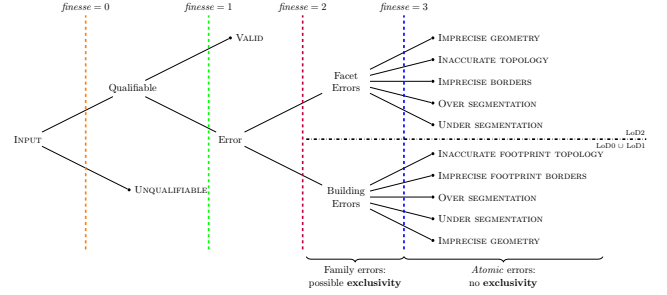


Fig. 2: The proposed taxonomy structure. In the VHR overhead case, only two families of errors are depicted. At *finesse* level 2, a hierarchy exists and **exclusivity** may prevail.

We propose the following *atomic* errors based on our comprehensive analysis on overhead aerial models.

#### - **Building errors** family (Figure 3i):

- *Under segmentation (BUS)*: two or more buildings are modeled as one;
- *Over segmentation (BOS)*: one building is subdivided into two or more buildings;
- *Imprecise footprint borders (BlmB)*: the building footprint borders are imprecise;
- *Inaccurate footprint topology (BlmT)*: the building footprint suffers from topological defects as missing inner courts or wrong primitive fitting;
- *Imprecise geometry (BIG)*: inaccurate building geometric estimation. In case  $\text{eLoD} > \text{LoD-0} \cup \text{LoD-1}$ , this error is not reported (redundant with errors below);

#### - **Facet errors** family (Figure 3ii):

- *Under segmentation (FUS)*: two or more facets are modeled as one;
- *Over segmentation (FOS)*: one facet is subdivided into two or more facets;
- *Imprecise borders (FIB)*: the facet borders are imprecise;
- *Inaccurate topology (FIT)*: the facet suffers from topological defects like a wrong primitive fitting;
- *Imprecise geometry (FIG)*: inaccurate facet geometric estimation.

#### B. Feature baseline

We propose a simple handcrafted set. We limit ourself to the validation of the semantic approach and taxonomy validation. We rely on three modalities. The first is the input 3D model. We extract statistics (maximum, minimum, mean, median and standard deviation) over attributes that describe the geometry of the building facets: number of nodes, area, angle between normals, distance between centroids and circumference. We add height attributes. These are represented by the histogram of the discrepancy between the rasterized model height map and the DSM at the same spatial resolution. It is chosen to be lower than building dimensions but higher than the modeling inherent planimetric uncertainty. Last comes optical features, computed by back-projecting the model on corresponding VHR orthoimage. For each facet edge, we compute

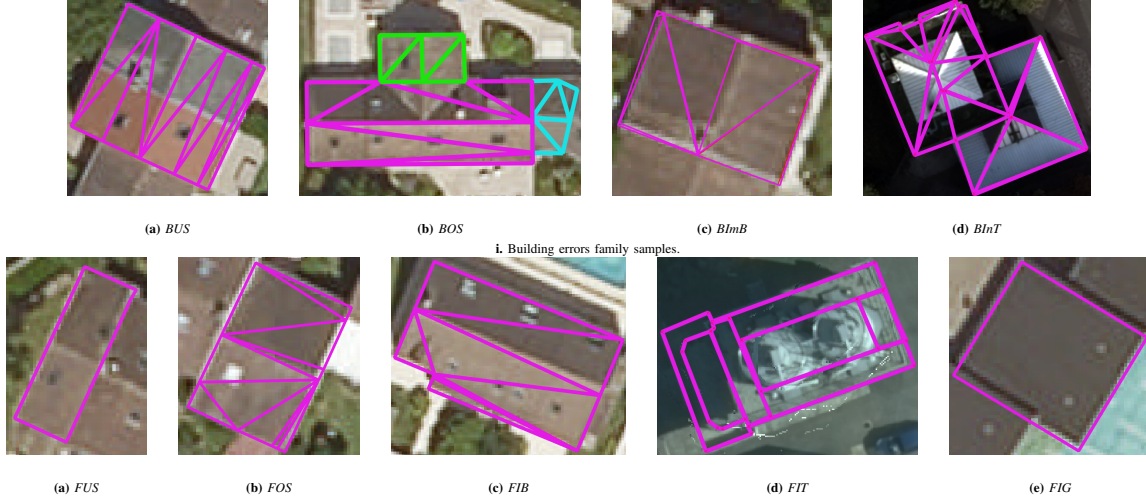


Fig. 3: Illustration of various errors of our taxonomy. One can see that geometric, spectral and height information are required for an accurate detection of all kinds of errors.

a correlation histogram to image edges (cosine similarity between local gradients and the edge normal). This histogram is summed over all facet edges and then over all projected facets. Histograms concentrated near the value 1 means facet edges correspond to genuine ones in the VHR image. Each modality provides 20 features.

### C. Classification process

The labels are organized into different classification problems. To account for the desired modularity and flexibility, a random forest classifier is chosen (1,000 trees, maximal depth of 4). The depth is kept shallow in order to avoid overfitting while the large number of estimators deals with the large feature space variability. The *Multi-label* setting is managed using a One vs All strategy on top of our classifier.

## IV. EXPERIMENTS

We evaluate our approach on a dataset drawn from three French cities: Elancourt, Nantes and one district of Paris (Paris-13). Elancourt exhibits a high diversity of building types: residential areas (bi-level buildings) and industrial sections (flat roofs). Nantes represents a denser urban setting with lower diversity. In Paris-13, high towers coexist with Haussmann style buildings. Input 3D models were obtained using the algorithm described in [3], out of existing footprints and an aerial VHR multi-view DSM. They were annotated according to the atomic errors list provided by our taxonomy. More details are reported in Table I.

|           | Elancourt | Nantes | Paris-13 |
|-----------|-----------|--------|----------|
| # samples | 2009      | 748    | 478      |
| DSM res.  | 6 cm      | 10 cm  | 10 cm    |
| Img. res. | 6 cm      | 10 cm  | 10 cm    |

TABLE I: Dataset details.

Labels are extracted from the taxonomy at *finesse* level of 3. In Table II, we report results over Elancourt using all four possible feature configurations performing a 10-fold cross validation. In Table III, we experimented training over buildings from one urban zone and testing on another.

From Table II, we can see that well represented errors are correctly detected. In fact, *BOS* (resp. *FOS*, *FIB*), with a 66.8% (resp. 64.16%, 59.08%) occurrence ratio, attains, in average across all configurations, a high recall ratio of 91.93% (resp. 98.93%, 79.68%). In contrast, other *atomic* errors which are not frequently found in the dataset (a~10% of the total) are naturally harder to detect. The ablation study reveals that, for most errors, geometric features yield sensibly the same results, except for *BUS* (where height and/or image based features add around 10% in recall while loosing at most 5% in precision), and for *BImB* (as image based features add around 4% in recall without a loss in precision). In Paris-13 and Nantes buildings are often more adjacent than Elancourt, this perfectly explains why classifiers trained over Paris/Nantes are better, regarding Building *atomic* errors, in recall and precision than the ones trained over Elancourt and tested on them. It achieves even close or better recall ratio over *BImB* if compared to classifiers trained on Elancourt itself. However, Facet *atomic* errors are much more easily detected if the classifiers are trained over Elancourt and tested on the other cities. It is also mostly stable compared to evaluation on Elancourt itself and even yields much better results for *FIB* and *FIG*. This can be explained by the fact that Haussmann style buildings and high flat towers offer less diverse facet geometries compared to bi-level hipped roofs, industrial buildings and other residential features that are present in the Elancourt scene.

Qualitative assessment is also performed in order to illustrate some failure cases. In the leftmost example of Figure 4, the similarity of the building outline to over segmented build-

|             | Geometry     |              | Geom. $\cup$ Height |              | Geom. $\cup$ Image |              | All          |              |
|-------------|--------------|--------------|---------------------|--------------|--------------------|--------------|--------------|--------------|
|             | Recall       | Prec.        | Recall              | Prec.        | Recall             | Prec.        | Recall       | Prec.        |
| <i>BOS</i>  | <b>93.96</b> | 76.15        | 91.43               | <b>77.76</b> | 91.51              | 76.08        | 90.83        | 76.14        |
| <i>BUS</i>  | 32.98        | <b>76.47</b> | <b>41.86</b>        | 75.57        | 40.38              | 71.00        | 39.32        | 71.81        |
| <i>BlmB</i> | 12.32        | 67.57        | 12.81               | <b>68.42</b> | 16.26              | 67.35        | <b>16.75</b> | 68.0         |
| <i>BlmT</i> | <b>25.25</b> | 92.59        | 20.20               | 90.91        | 20.20              | <b>95.24</b> | 11.11        | 91.67        |
| <i>FOS</i>  | 98.91        | 99.07        | 98.91               | <b>99.30</b> | <b>98.99</b>       | 98.84        | 98.91        | 98.84        |
| <i>FUS</i>  | <b>1.90</b>  | 54.55        | 0.63                | <b>66.67</b> | 1.61               | 50           | 1.27         | <b>66.67</b> |
| <i>FIB</i>  | <b>9.17</b>  | 87.5         | 0                   | —            | 8.30               | 82.61        | 7.42         | <b>100</b>   |
| <i>FIT</i>  | <b>6.67</b>  | <b>100</b>   | 8.73                | 95.24        | 3.33               | <b>100</b>   | 3.33         | <b>100</b>   |
| <i>FIG</i>  | <b>80.54</b> | 73.14        | 80.45               | <b>72.62</b> | 78.69              | 72.12        | 79.02        | 71.82        |

TABLE II: Ablation study (in %) at *finesse* = 3 on Elancourt.

|             | Elancourt $\rightarrow$ Nantes |       | Elancourt $\rightarrow$ Paris-13 |       | Nantes $\rightarrow$ Elancourt |       | Paris-13 $\rightarrow$ Elancourt |       |
|-------------|--------------------------------|-------|----------------------------------|-------|--------------------------------|-------|----------------------------------|-------|
|             | Recall                         | Prec. | Recall                           | Prec. | Recall                         | Prec. | Recall                           | Prec. |
| <i>BOS</i>  | 100                            | 42.05 | 78.71                            | 42.97 | 86.47                          | 64.99 | 82.14                            | 69.59 |
| <i>BUS</i>  | 19.12                          | 54.17 | 4.76                             | 27.27 | 23.75                          | 57.47 | 25.0                             | 50.49 |
| <i>BlmB</i> | 0                              | —     | 1.85                             | 33.33 | 15.65                          | 46.75 | 19.89                            | 48.05 |
| <i>BlmT</i> | 40.71                          | 14.24 | 0                                | —     | 4.70                           | 100   | 4.44                             | 100   |
| <i>FOS</i>  | 100                            | 69.38 | 98.80                            | 96.47 | 98.76                          | 98.92 | 99.06                            | 97.14 |
| <i>FUS</i>  | 15.71                          | 56.90 | 20.73                            | 91.94 | 1.68                           | 77.78 | 3.53                             | 89.47 |
| <i>FIB</i>  | 83.54                          | 43.08 | 39.58                            | 64.04 | 23.75                          | 73.96 | 21.86                            | 70.11 |
| <i>FIT</i>  | 9.09                           | 33.33 | 0                                | 0     | 3.57                           | 100   | 4.34                             | 33.33 |
| <i>FIG</i>  | 100                            | 64.45 | 90.86                            | 86.14 | 86.35                          | 67.97 | 82.38                            | 73.08 |

TABLE III: Transferability study (in %) at *finesse* = 3 over different urban scenes using all features. Compared to previous results in Table II, ■: a stable or better result, ■: a small drop and ■: a significant decrease in accuracy.

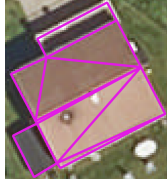
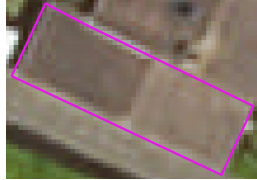
|                                                                                   |                                                                                   |                                                                                    |                                                                                     |      |       |            |      |       |             |      |       |
|-----------------------------------------------------------------------------------|-----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|------|-------|------------|------|-------|-------------|------|-------|
|  |  |  |  |      |       |            |      |       |             |      |       |
| Errors                                                                            | G.T.                                                                              | Pred.                                                                              | Errors                                                                              | G.T. | Pred. | Errors     | G.T. | Pred. | Errors      | G.T. | Pred. |
| <i>BOS</i>                                                                        | ✗                                                                                 | ✓                                                                                  | <i>BUS</i>                                                                          | ✗    | ✓     | <i>BOS</i> | ✓    | ✓     | <i>BOS</i>  | ✓    | ✗     |
| Valid                                                                             | ✓                                                                                 | ✗                                                                                  | <i>FIG</i>                                                                          | ✓    | ✗     | <i>FUS</i> | ✓    | ✗     | <i>FOS</i>  | ✓    | ✗     |
|                                                                                   |                                                                                   |                                                                                    | <i>FOS</i>                                                                          | ✓    | ✗     |            |      |       | <i>BUS</i>  | ✓    | ✓     |
|                                                                                   |                                                                                   |                                                                                    |                                                                                     |      |       |            |      |       | <i>BlmB</i> | ✓    | ✓     |

Fig. 4: Predicted (Pred.) errors compared to ground truth (G.T.) labels are illustrated in some pathological cases. Knowing how each error is represented in the dataset helps interpreting misclassifications.

ings cases induces an overdetection. Sometimes, the VHR of the data does not suffice to detect all errors that are visually identifiable. In general, these samples help us devise more robust features so as to alleviate these issues. Dataset enrichment could be another option which provides more instances of underrepresented errors. In the end, we can also add the human in the loop through a manual interactive evaluation/classification mechanism which could adapt well to user-specific needs.

## V. CONCLUSION

We proposed a framework to semantically evaluate 3D building models. Errors were hierarchically organized into a flexible and parametrizable taxonomy, handling the large diversity of urban environments and end users' requirements. Model quality was predicted using a supervised classifier using three modalities. Although being mitigated over under-represented errors, results are satisfactory in the well balanced cases. As a next step, more structurally aware features (e.g., based on graph comparison) will be proposed, so as to be applied on a richer and more diverse dataset (potentially involving data augmentation) under a deep-based framework.

## REFERENCES

- [1] P. Musialski, P. Wonka, D. G. Aliaga, M. Wimmer, L. van Gool, and W. Purgathofer, "A survey of urban reconstruction," *EUROGRAPHICS 2012 State of the Art Reports*, vol. XX, pp. 1–28, 2012.
- [2] T. H. Kolbe, G. Gröger, and L. Plümer, "CityGML: Interoperable access to 3D city models," in *Geo-information for disaster management*. Springer, 2005, pp. 883–899.
- [3] M. Durupt and F. Taillandier, "Automatic building reconstruction from a Digital Elevation Model and cadastral data: an operational approach," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 36, no. 3, pp. 142–147, 2006.
- [4] H. Kaartinen *et al.*, "Accuracy of 3d city models: Eurosdrc comparison," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 36, no. 3/W19, pp. 227–232, 2005.
- [5] H. Ledoux, "val3dity: validation of 3D GIS primitives according to the international standards," *Open Geospatial Data, Software and Standards*, vol. 3, no. 1, p. 1, 2018.
- [6] C. Zeng, S. Member, T. Zhao, and J. Wang, "A multicriteria evaluation method for 3D building reconstruction," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 9, pp. 1619–1623, 2014.
- [7] L. Boudet, N. Paparoditis, F. Jung, G. Martinoty, and M. Pierrot-Deseilligny, "A supervised classification approach towards quality self-diagnosis of 3D building models using digital aerial imagery," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 36, no. 3, pp. 136–141, 2006.
- [8] J.-C. Michelin, J. Tierny, F. Tupin, C. Mallet, and N. Paparoditis, "Quality evaluation of 3D city building models with automatic error diagnosis," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XL-7/W2, pp. 161–166, 2013.