Estimating Completeness of VGI Datasets by Analyzing Community Activity Over Time Periods

Simon Gröchenig, Richard Brunauer and Karl Rehrl

Abstract Due to the dynamic nature and heterogeneity of Volunteered Geographic Information (VGI) datasets a crucial question isu concerned with geographic data quality. Among others, one of the main quality categories addresses data completeness. Most of the previous work tackles this question by comparing VGI datasets to external reference datasets. Although such comparisons give valuable insights, questions about the quality of the external dataset and syntactic as well as semantic differences arise. This work proposes a novel approach for internal estimation of regional data completeness of VGI datasets by analyzing the changes in community activity over time periods. It builds on empirical evidence that completeness of selected feature classes in distinct geographical regions may only be achieved when community activity in the selected region runs through a well-defined sequence of activity stages beginning at the start stage, continuing with some years of growth and finally reaching saturation. For the retrospective calculation of activity stages, the annual shares of new features in combination with empirically founded heuristic rules for stage transitions are used. As a proof-of-concept the approach is applied to the OpenStreetMap History dataset by analyzing activity stages for 12 representative metropolitan areas. Results give empirical evidence that reaching the saturation stage is an adequate indication for a certain degree of data completeness in the selected regions. Results also show similarities and differences of community activity in the different cities, revealing that community activity stages follow similar rules but with significant temporal variances.

S. Gröchenig · R. Brunauer (✉) · K. Rehrl
Salzburg Research Forschungsgesellschaft mbH, Jakob-Haringer-Straße 5,
5020 Salzburg, Austria
e-mail: richard.brunauer@salzburgresearch.at

S. Gröchenig
e-mail: simon.groechenig@salzburgresearch.at

K. Rehrl
e-mail: karl.rehrl@salzburgresearch.at

J. Huerta et al. (eds.), *Connecting a Digital Europe Through Location and Place*,
Lecture Notes in Geoinformation and Cartography, DOI: 10.1007/978-3-319-03611-3_1,

1 Introduction

Volunteered Geographic Information (VGI) denotes one of the most promising and interesting developments in the field of geographic information science. Since the coining of the term by Goodchild (2007) researchers all over the world have started to scientifically investigate the phenomenon. One of the most crucial questions in VGI research is concerned with the assessment of geographic data quality of VGI datasets (ISO 2011; Goodchild and Li 2012). One of the outstanding categories of geographic data quality addresses completeness. Although completeness estimations of geographic datasets are not new, the VGI movement raises some new aspects such as inherent heterogeneity, high regional differences or frequent changes. In previous work researchers have addressed the assessment of data completeness in VGI datasets with well-known approaches like comparisons with external reference datasets (Haklay et al. 2010; Mondzech and Sester 2011; Zielstra and Hochmair 2011). Due to the success and open license of OpenStreetMap (OSM) (Haklay and Weber 2008) the project has been focus of most previous studies. The number of features, total lengths of linear features or the overlapping area of buffered features are compared. Although significant progress has been achieved, comparisons with external reference datasets have certain disadvantages such as the incertitude concerning completeness of the reference datasets, the missing of global availability or legal restrictions as well as high fees (Hecht et al. 2013). To overcome these disadvantages this work introduces a novel approach aiming at internal evaluation of data completeness. The presented approach analyzes the community activity over time periods in order to determine whether a certain level of completeness has been reached in a selected region. For estimating the completeness level in a region the approach derives the three activity stages *Start*, *Growth* and *Saturation* from the annual increase of geographic features being mapped by volunteering community members. The measure for completeness estimation is based on the hypotheses that completeness in a region can only be achieved when community activity passes a well-defined sequence of activity stages.

The remainder of this chapter is organized as follows: The next section discusses related work. It is followed by a section on the theoretical aspects of assessing completeness of VGI datasets. Section 4 outlines the novel approach for internal estimation of data completeness. Section 5 introduces the dataset used for proof-of-concept evaluation. Section 6 presents and discusses results and finally, Sect. 7 concludes the work.

2 Related Work

Goodchild and Li (2012) propose three approaches for quality assurance of VGI datasets: the crowdsourcing approach (i) relies on the community to check each other's contribution, the social approach (ii) gives people the responsibility of moderating the mapping process and the geographic approach (iii) deals with correctness

of spatial data. For measuring data quality of already mapped features related work considers ISO 19157 (2011) where standardized quality measures for geographic information are defined. Related work mainly addresses the quality categories *completeness* and *positional accuracy* for the most prominent open VGI dataset OSM.

Haklay (2010) compared the OSM street network (motorways, A- and B-roads) of London with the federal dataset provided by Ordnance Survey. He concluded that on average 80 % of the streets are already mapped. Neis et al. (2012) compared the OSM street network of Germany with the commercial data provided by TomTom. They showed that OSM has a longer street network for pedestrians while TomTom is more detailed at rural street networks for cars. Moreover, authors revealed that urban street networks developed earlier than rural ones. Similarly, Zielstra and Hochmair (2011) compared the street network in selected cities in Germany and in the US with three reference datasets, namely Tiger, NAVTEQ and TomTom. Girres and Touya (2010) conducted a similar study for French roads, rivers and lakes. They determined a relative completeness of 45 % for roads, 83 % for lakes and 8 % for rivers compared to the French IGN dataset. Hecht et al. (2013) compared the OSM buildings with the ALKIS/ATKIS datasets for selected regions in Germany and concluded that less than 30 % of all buildings have been mapped.

While previous approaches pursue external data quality measures, the following studies focus on internal measures without relying on reference datasets. One of the first internal quality assessments was done by Mooney et al. (2010) who examined the geometry of polygons. Neis et al. (2013) analyzed the development of OSM data in 12 metropolitan areas distributed all over the world. According to their analysis of active users, European cities show a more active OSM community. Furthermore, authors analyzed the creation date and latest update of all features. They found that more than 20 % of all features have been created in 2012 and used this as indicator, that the dataset is not complete, yet. Corcoran et al. (2013) proofed that the growth of OSM street networks follows the development pattern of street networks in the real world defined by Strano et al. (2012). This pattern describes that the exploration phase (when new areas are mapped) is followed by a densification phase (when more details are added). Barron et al. (2014) developed a tool to analyze 25 indicators for assessing OSM data quality. Arsanjani et al. (2013) simulate the OSM mapping development for upcoming years based on development in previous years.

From examining previous work it can be concluded that only few studies address internal completeness measures of VGI datasets. To the knowledge of the authors there is no approach analyzing the development of the community activity over time periods for estimating regional data completeness.

3 Estimating Completeness of VGI Datasets

The International Organization for Standardization defines in ISO 19157 "Geographic information—Data quality" (ISO 2011) five data quality categories for geographic information, namely *completeness*, *logical consistency*, *positional*

accuracy, *thematic accuracy* and *temporal quality* where completeness, which is addressed in this chapter, is defined as:

> […] the presence and absence of features, their attributes and relationships. It consists of two data quality elements: commission—excess data present in a dataset; and omission—data absent from a dataset.

The completeness of a dataset depends on the presence of features in the dataset and on the correspondence between these features and the objects or properties in the real world. The measure does not depend on the positional accuracy or on the level of detail of the features. Completeness is a property of a geographical dataset and restricted to a geographical area and a purpose. The purpose defines the set of feature classes which are investigated. Hence we define completeness as:

> The *completeness measure* of the geographical dataset D, where D is defined by geographical region R and for purpose P, depends on the degree of correspondence between the existence of objects and properties in the real world and the presence of their representing features in dataset D.

However, the degree of correspondence (i) cannot be measured directly and the value (ii) cannot be calculated from the geographical dataset alone. Thus, completeness is commonly estimated by comparing two geographical datasets where the reference dataset is used instead of the real world. The comparisons of datasets with reference datasets or with the real world are so-called "external approaches" of quality evaluations, while internal approaches estimate data quality by calculating quality parameters from the dataset itself (ISO 2011). Internal approaches have to use well-defined rules to derive completeness indicators. The adequacy of such rules has to be proofed empirically. For VGI, the following three rules for estimating data completeness can be applied:

1. **Community activity and contributions** One possibility to estimate internal data completeness is to conduct an analysis of community contributions to VGI datasets (Neis et al. 2012; Steinmann et al. 2013a, b). Characteristic of VGI datasets are frequently appearing, disappearing or changing features. Since these changes are assigned to their contributors the current development of mapping activity of the community may always be treated as indicator for data completeness (e.g. Neis et al. (2013)).
2. **Hierarchical relationships between feature classes** In VGI feature classes are typically mapped according to their importance and appearance. For example, motorways are usually mapped before lower-level streets (Neis et al. 2012). An approach for estimating completeness may consider such hierarchical structures (e.g. Corcoran et al. (2013)). Thus, the temporal appearance of feature classes or feature class combinations may be used as completeness indicator.
3. **Relations between neighboring, sub- and super-regions** Completeness assessments have not to be treated as regionally isolated tasks. For example, it seems obvious that complete regions are more likely to be in a cluster of complete neighbors or that they contain at least complete sub-regions (Arsanjani et al. 2013). Together with the results of other rules relationships between spatially close and equally developed regions could be used as characteristic completeness indicator.

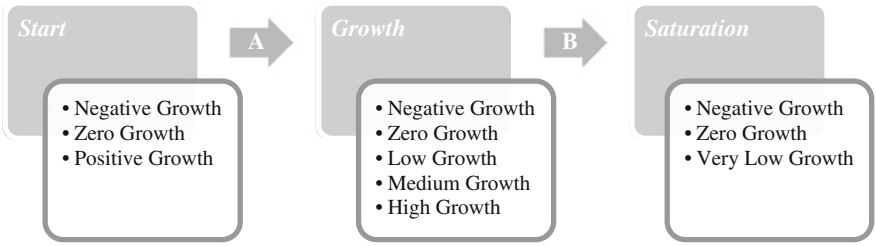


Fig. 1 Activity stages for analyzing the completeness of VGI datasets

In this work we outline an approach for completeness estimation which is a combination of rule types (1)–(3). As a first step, the growth rates of features in a dataset are analyzed to derive annual stages of community activity (rule type (1)). The activity stages represent an empirically determined mapping progress where the last stage is supposed to be a proper candidate indicator for completeness. Additionally, a detailed analysis of community activities is conducted by regarding rules from type (2) and (3). The results are used to gather additional evidence that completeness in a certain region has been reached or is near to be reached.

4 Deriving Community Activity Stages

Activity stages describe the contributors' activity by analyzing the annual changes to features in a dataset. The stages describe an ideally unidirectional development of the activities: at the start of a community activity only a few contributors are contributing to the dataset, afterwards more contributors are joining the activity and start contributing data before the mapping activity ceases since a certain level of data completeness has been reached. The development of these activities is described with the three stages *Start*, *Growth* and *Saturation* (Fig. 1). Within a certain stage community activity may change to more detailed sub-stages. The transitions between main stages follow distinct rules and are typically unidirectional. For the presented analysis the activity stage *Saturation* is the most relevant one. It occurs in the final years of a development in case that no more new (or just few) features are created.

The definition of stage transitions and sub-stage classifications is based on a growth value. For time interval i, region r and feature class f the growth value g is defined as the difference between the number of created features c and the number of deleted features d

$$g\,(i, r, f) := c\,(i, r, f) - d\,(i, r, f) \tag{1}$$

The progress value p is defined for time interval i, region r and feature class f as the fraction of the growth value from the overall growth value over the whole analyzed time interval $I\,(i \subseteq I)$

Table 1 Transition rules for sub-stages

Activity stage	Condition	Sub-stage
For all stages	$p(i, r, f, I) < 0$	Negative growth
	$p(i, r, f, I) = 0$	Zero growth
Start	$0 < p(i, r, f, I)$	Positive growth
Growth	$0 < p(i, r, f, I) \leq 0.25$	Low growth
	$0.25 < p(i, r, f, I) \leq 0.75$	Medium growth
	$0.75 < p(i, r, f, I)$	High growth
Saturation	$0 < p(i, r, f, I) \leq 0.03$	Very low growth

$$p(i, r, f, I) := g(i, r, f) / g(I, r, f) \tag{2}$$

A progress value of 0.36 indicates that 36 % of all features have been created in the respective time period. Transition rules between sub-stages are shown in Table 1.

In Fig. 1 the rules for the unidirectional transitions *A* and *B* between the activity stages are empirically defined as heuristic rules. For transition *A* from *Start* to *Growth*, two or more active contributors within a distinct region are required. Transition *B* from *Growth* to *Saturation* requires the progress within a time period to be very low (less than 3 %), whereas the cumulated progress value is greater than 0.97 and the number of years with active contributions is greater than two. Due to the retrospective calculation of growth values the resulting activity stages and sub-stages are subject to change. Since community effort is continuously changing, significant annual growth may occur although there was only minor growth during the previous years. It should be noted that any re-evaluation of the dataset with additional data may result in other activity stages for the previous years. It should also be noted that transition rules, although being derived from empirical evidence, should be treated as 'subject to change' since additional analyses could reveal the necessity of adjustments.

5 OSM History as Evaluation Dataset

To evaluate the proposed measure for data completeness, it is applied to the historic changes of the OSM data (OSM History). The OSM History has been selected since it includes all versions of all features starting from 2006 until the current date (for this analysis the file from 5th Feb. 2013 has been used). Since the calculation of activity stages is based on the definition of the growth value 4.1 and progress value 4.2 the historic data has to be prepared. Data preparation is based on the algorithm proposed in Rehrl et al. (2012) and results in a list of annually aggregated growth shares based on the total number of created and deleted features for the selected year. In addition to growth shares, the total number of active contributors is calculated for each year. Annual growths are considered as well-suited temporal units by avoiding seasonal variability. The geographic scope of the analysis has been set to the same

12 metropolitan areas (equal delimitation) as proposed in Neis et al. (2013). Besides fostering comparability of results the selected areas are considered well-suited due to worldwide distribution, cultural diversity and homogeneous settlement structure with a large number of geographic features and feature classes. Moreover, it has been previously found that urban communities are commonly more developed and more active (Neis et al. 2012). For proving the results of the 12 metropolitan areas, the three Austrian cities Vienna, Linz and Salzburg have been added. At least the mapping of the street network has been estimated "complete" by the local OSM community (OSM Wiki 2013b) and the authors' local knowledge confirms this estimation. Thus, the results for the Austrian cities are used as ground-truth for the proposed completeness measure.

Beside geographic scopes, completeness measures have to be focused on different feature classes. While OSM does not follow strict rules for classifying features, the proposed keys and values in the OSM Wiki may be used for selecting feature classes (OSM Wiki 2013a). As previously found, the feature classes denoted by the keys *highway* and *building* are significantly more developed in comparison to all other classes (Steinmann et al. 2013a). Due the high development it may be assumed that both classes have passed several years of mapping activity in all of the selected regions.

In OSM the key *highway* comprises all kinds of features related to the street network. This includes motorways, roads, residential streets, tracks, paths and footways. The highway key is also used for point features like traffic lights, turning points or pedestrian crossings. Due to the heterogeneous nature of the feature class it is suggested to analyze sub-classes separately.

The key *building* is used for mapping each kind of buildings. The value specifies the type of building (e.g. residential buildings, hotels or churches). In contrast to highways, the building class is homogenously structured and thus may be analyzed as a whole. In addition to the footprint and the building type, additional information such as addresses may be attributed to buildings. Since address information is typically mapped after building footprints, a separate analysis is suggested.

For the evaluation of the completeness measure, four different feature classes are selected: (i) the class *street* subsumes the OSM *highway* sub-classes *primary*, *secondary*, *tertiary*, *living_street*, *residential* and *unclassified*, (ii) the class *path* subsumes the sub-classes *path*, *footway*, *cycleway* and *steps*, (iii) the class *building* regards all features having the key *building* and finally (iv) the class *house number* regards all features having the key *addr:housenumber*. While the classes *street* and *building* are mainly used for calculating community activity, the classes *path* and *house number* are used as additional indicators to estimate the level of completeness.

6 Results and Discussion

This section presents selected evaluation results and discusses the results in the context of the following criteria: (i) impacts of different spatial resolutions, (ii) time series of activity stages to highlight the transitions from *Start* to *Growth* to *Saturation*,

Fig. 2 Activity stages for London using three spatial resolutions (feature class: street; year: 2012): **a** shows the activity stage for Greater London. **b** shows activity stages for the 32 boroughs plus the City of London. **c** shows activity stages as hexagon grid consisting of cells with a diameter of 5 km; administrative boundary (*black line*); metropolitan area (*white line*) defined by Neis et al. (2013)

(iii) comparisons of activity stages for the selected metropolitan areas for one year, (iv) comparisons of activity stages between selected feature classes and (v) comparisons of spatial activity stage patterns of the last year.

Figure 2 shows the impact of different spatial resolutions on the calculation of activity stages for the London metropolitan area based on the same delimitation used by Neis et al. (2013). Firstly, the algorithm is applied to Greater London resulting in one conflated activity stage, in the middle the 32 London Boroughs plus the City of London are analyzed separately resulting in different activity stages and on the right the metropolitan area is subdivided by a hexagon grid with a cell diameter of five kilometers. According to Hagenauer and Helbich (2012) the shapes of hexagon grids follow urban patterns best. While the former two resolutions are bound to administrative boundaries the third one ignores boundaries. The benefit of using a grid resolution can be found in the worldwide applicability as well as in the comparability of different world regions. Analyses based on administrative boundaries cannot be compared due to variances in size and shape. For example, Great Britain is subdivided by different administrative structures with totally different sizes. Moreover it has been previously found that coarse-grained spatial resolutions with larger regions conflate individual results, which could get apparent with more fine-grained resolutions (Haklay et al. 2010). The London example from Fig. 2 confirms this finding for activity stages as the spatial resolution of Greater London conflates the different activity stages of the boroughs. However, activity stages for unpopulated areas should be specifically addressed due to lower mapping activity. Indeed it should be noted that larger evaluation units (administrative boundaries) may be useful for more general analyses. Table 2 summarizes advantages and disadvantages of the three proposed spatial resolutions with emphasis on analyzing community activity. The remainder of this work builds on the hexagon approach.

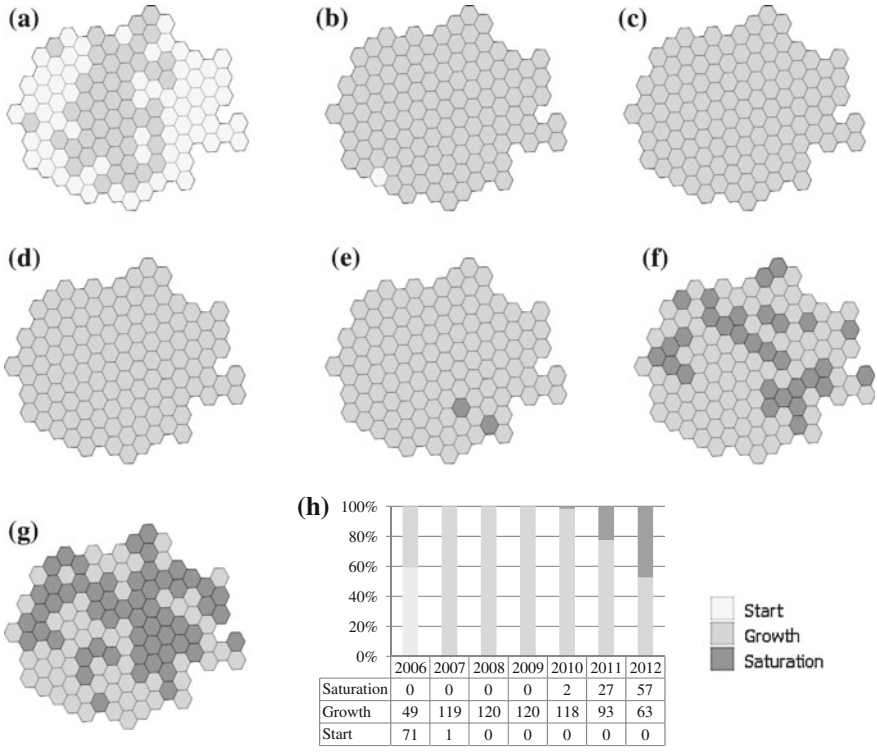
Figure 3 shows the sequential changes of community activity stages (see rule (1) in Sect. 3) for the feature class *street* in the metropolitan area of London during the years 2006–2012 using hexagon cells with 5 km diameter. The annual results shown in the hexagon maps are summarized in the bottom right diagram. While in the year

Table 2 Advantages and disadvantages of different spatial resolutions

	Greater London	Boroughs + City	Hexagons
Advantage	Fewer test regions; overview analysis	Residential areas are considered, no areas without population or infrastructure	All polygons have the same size and emphasize; world-wide comparability; detailed analysis of homogenous topographies (e.g. big cities)
Disadvantage	Places with different activity stages are conflated; no detailed conclusions are possible	Places with different activity stages are conflated; detailed conclusions only with additional contextual knowledge; large polygons are more emphasized in visualization	Areas with low contribution level; especially unpopulated areas; hexagons do not fit administrative boundaries

2006 71 cells are still in *Start*, in 2007 119 out of 120 cells have proceeded to *Growth* which is an indication for rising community activity in all parts of London. Since London has been the incubator city of the OSM project, activity stages are temporally ahead in comparison to other cities. Community activity most likely starts in the city center and moves towards the suburbs subsequently (see Cairo in Fig. 5 and Buenos Aires in Fig. 6). In 2010, the first two cells reached *Saturation*. In 2012, a majority of cells has reached *Saturation* which can be interpreted as indication that a certain level of completeness has been achieved.

To address the question whether saturated cells are also complete cells the next analysis regards the hierarchical structure of the feature classes (see rule (2) in Sect. 3). Tables 3 and 4 compare the activity stages of 12 metropolitan areas and three Austrian reference cities for the year 2012. While Table 3 shows the results for feature classes *street* and *path*, Table 4 has its focus on comparing the classes *building* and *house number*. The tables show (i) the number of hexagon cells per metropolitan area or city, (ii) the absolute numbers of cells which are in *Start* and *Saturation*, respectively, and charts showing the relative share of the three activity stages and (iii) ratios between the numbers of created features between the related feature classes. Results in Table 3 emphasize that streets are mapped before paths, while Table 4 indicates that buildings are mapped before house numbers. In case of streets in Berlin, 34 of 85 hexagons have already reached *Saturation* by the end of 2012 while for the paths only 14 hexagons have achieved the final stage. Results indicate that cities with similar activity patterns exist. In Table 3, the cities Berlin,



	2006	2007	2008	2009	2010	2011	2012
Saturation	0	0	0	0	2	27	57
Growth	49	119	120	120	118	93	63
Start	71	1	0	0	0	0	0

Fig. 3 Activity stages (2006–2012) for feature class *street* for the metropolitan area of London showing the mapping progress. **a** 2006, **b** 2007, **c** 2008, **d** 2009, **e** 2010, **f** 2011, **g** 2012, **h** Shares of activity stages of the 120 hexagon cells of London as time series from 2006–2012

London, Los Angeles, Moscow, Paris and Sydney have a faster mapping progress than the other cities. These cities also have the most advanced ratio towards mapping paths which is an indication for a high level of street completeness. Contrary, Johannesburg has a very low ratio and a high proportion of saturated cells, which is an indicator for a temporary inactive community than for completeness. For buildings and house numbers, the cities Buenos Aires, Cairo and Istanbul have the slowest activity progress. Due to Japan's different address scheme, only a very low number of house numbers has been mapped in Osaka which results in an adverse ratio (similar for Seoul). This finding gives indication that slow progress can also be the result of cultural variations. In case of Paris it has to be stressed that the city benefited from a major building import in 2010. In both tables the results for the three Austrian reference cities Linz, Salzburg and Vienna reveal the supposed advanced mapping progress with Linz as the most saturated city for all feature classes which reflects own and community observations. The overall results confirm previous results from Neis et al. (2013) that European cities have a more active OSM community compared to other cities.

Table 3 Comparisons between 12 metropolitan areas + 3 reference cities based on activity stages of the year 2012 for the feature classes street and path

Highway 2012	Hex. #	Street			Path			Ratio s:p
		Start	Sat.	%	Start	Sat.	%	
Berlin	85	0	34		0	14		1 : 1.5
Buenos Aires	177	8	52		134	7		13.0 : 1
Cairo	105	21	34		93	3		19.5 : 1
Istanbul	110	19	21		91	1		18.7 : 1
Johannesburg	167	12	80		140	4		24.7 : 1
London	120	0	57		0	7		1.2 : 1
Los Angeles	369	6	157		186	23		4.9 : 1
Moscow	300	1	113		30	14		1 : 1.9
Osaka	224	18	11		80	3		14.6 : 1
Paris	188	0	44		10	12		1.5 : 1
Seoul	147	39	15		112	3		12.4 : 1
Sydney	126	1	4		12	21		8.0 : 1
Linz	14	0	9		0	5		1 : 7.7
Salzburg	13	0	2		0	0		1 : 1.5
Vienna	41	0	9		0	3		1 : 1.9

Start Growth Saturation

The ratio s:p is the proportion between the number of created *street* (s) and *path* features (p)

A more detailed comparison demonstrates the shift in mapping progress between different hierarchically structured features. Figure 4 shows two time series for London and Paris. For London, the time series compares the progress of feature classes *street* and *path*, while for Paris, the progress of class *building* is compared to the progress of class *house number*. Both cities have been selected due to their advanced progress for the respective feature classes (see Tables 3 and 4). Figure 4 illustrates the shares of activity stages based on the hexagons for the years between 2006 and 2012. The diagrams outline that a shift between the related feature classes is observable. Streets are mapped before paths while buildings are mapped before house numbers. A possible reason for that phenomenon is that the focus of mapping interests follows a hierarchical order being determined by hierarchical relationships between feature classes. For example, most building footprints are mapped before house numbers are added. Based on this observations it may be valid to assess saturated hexagons more likely as complete if the observed area followed the typical hierarchical mapping schema, too. This trend can also be observed in the hexagon maps visualized in Figs. 5 and 6 and confirms the findings of Corcoran et al. (2013) about the exploration and densification phase.

Table 4 Comparisons between 12 metropolitan areas + 3 reference cities based on activity stages of the year 2012 for the feature classes building and house number

Building 2012	Hex. #	Building			House number			Ratio b:hn
		Start	Sat.	%	Start	Sat.	%	
Berlin	85	1	18		4	3		1.6 : 1
Buenos Aires	177	126	0		160	1		3.3 : 1
Cairo	105	79	5		100	1		21.7 : 1
Istanbul	110	78	1		98	0		27.3 : 1
Johannesburg	167	93	1		166	0		2.4 : 1
London	120	8	13		56	3		3.0 : 1
Los Angeles	369	178	25		249	8		5.8 : 1
Moscow	300	26	39		119	7		3.9 : 1
Osaka	224	137	0		220	0		1657 : 1
Paris	188	11	70		83	8		7.6 : 1
Seoul	147	128	0		140	0		11.6 : 1
Sydney	126	55	9		101	0		3.1 : 1
Linz	14	0	6		0	4		4.3 : 1
Salzburg	13	0	0		1	1		2.1 : 1
Vienna	41	0	0		1	3		1.6 : 1

Start Growth Saturation

The ratio b:hn is the proportion between the number of created *building* (b) and *house number* features (hn)

Concerning relationships between neighboring cells (see rule (3) in Sect. 3), the last comparison outlines the activity stage for the year 2012 for cities with a fast and a slow mapping progress. Figure 5 compares the results of London and Cairo for the feature classes *street* and *path* using hexagon maps. While London depicts advanced progress with respect to the mapping of streets, path mapping is still predominately stuck in *Growth*. Similarly, street mapping activity in Cairo is more advanced than path mapping. In contrast to London, Cairo still has 20 % of all hexagon cells for the street class and 89 % of the cells for the path class in *Start*. The high percentage of cells in *Start* for both feature classes indicates a low level of completeness for Cairo.

Figure 6 shows differences in building and house number mapping for the metropolitan areas of Paris and Buenos Aires. As Table 4 indicates, Paris has a high proportion of cells in *Saturation* while Buenos Aires still has many cells in *Start*. Again, more progress has been identified for buildings compared to house numbers. Single and distributed *Saturation* hexagons, as for paths in London (Fig. 5) and house numbers in Paris (Fig. 6), should be treated carefully. Those hexagons can also indicate temporal inactivity at the beginning of the mapping progress.

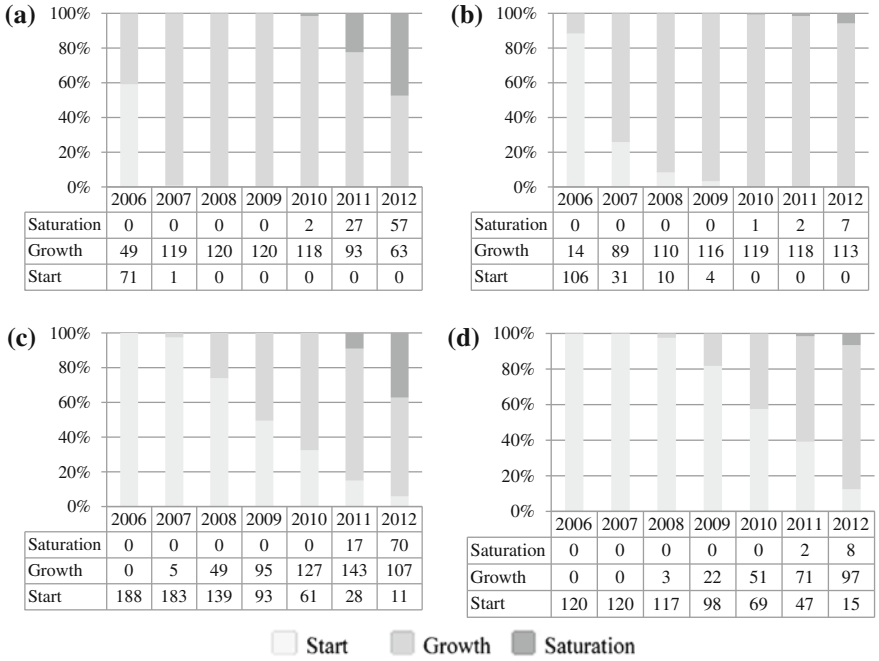


Fig. 4 Mapping progress based on activity stages (2006–2012) for London and Paris. **a** London—*street*, **b** London—*path*, **c** Paris—*building*, **d** Paris—*house number*

To summarize, it can be concluded that combining different rules for interpreting saturated cells leads to more accurate estimations. The presented examples have especially shown that (i) an appropriate selection of different spatial resolutions, (ii) the consideration of hierarchical structures between feature classes and (iii) the consideration of spatial distributions provide a proper analysis method for completeness.

However, the proposed approach is based on the simple hypothesis that when contributors cease to create features in a region, a sufficient level of completeness has been reached. A drawback of the method comes from the fact that low mapping activity can also be the result of non-ideal community developments (Suh et al. 2009). Thus, a critical evaluation of resulting activity stages is necessary for adequate estimations.

7 Conclusions and Outlook

In this chapter we proposed a new method for analyzing changes in VGI datasets to determine community activity stages in order to estimate regional completeness. The presented results show that local community activities provide sufficient information

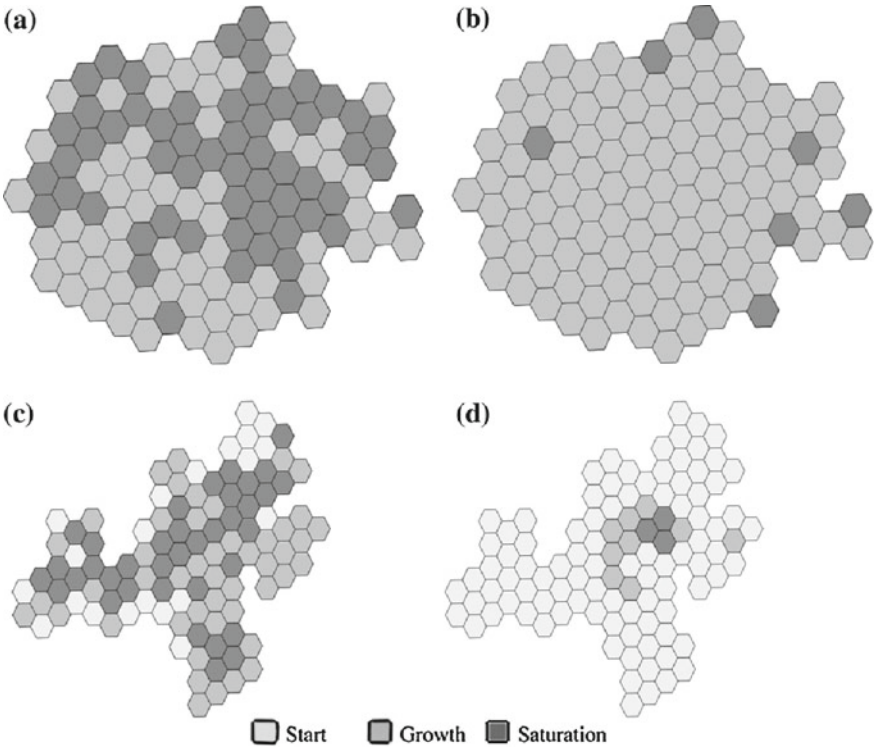


Fig. 5 Comparison of two different progress patterns in London and Cairo based on calculated activity stages for the year 2012 using the feature classes *street* and *path*. **a** London—*street*, **b** London—*path*, **c** Cairo—*street*, **d** Cairo—*path*

for assessing regions as complete. Several indicators show that regions with a low local community activity are estimated to be "regional complete", too.

The examples outlined above lead to plausible indications that the level of regional completeness can be derived from the temporal progress of community activity. Together with a detailed analysis of spatial distributions of activity stages a more accurate estimation between inactive and complete can be achieved. Furthermore, a selective analysis regarding the mapping progress of hierarchically structured feature classes, e.g. for streets and paths, facilitates the understanding and estimation of completeness. The proposed method can be easily adapted to different time periods, temporal resolutions, spatial resolutions and feature classes in order to provide deeper insights into the mapping progress of VGI communities.

To achieve reasonable results, the interpretation of activity stages for estimating completeness requires consideration of multiple aspects such as different hexagon sizes or related feature types. This applies especially for regions with a small or young VGI community or for regions with different cultural or topographic characteristics.

An open issue for further analyses would be to investigate more diverse geographic regions on applicability and generalization of the approach. For example

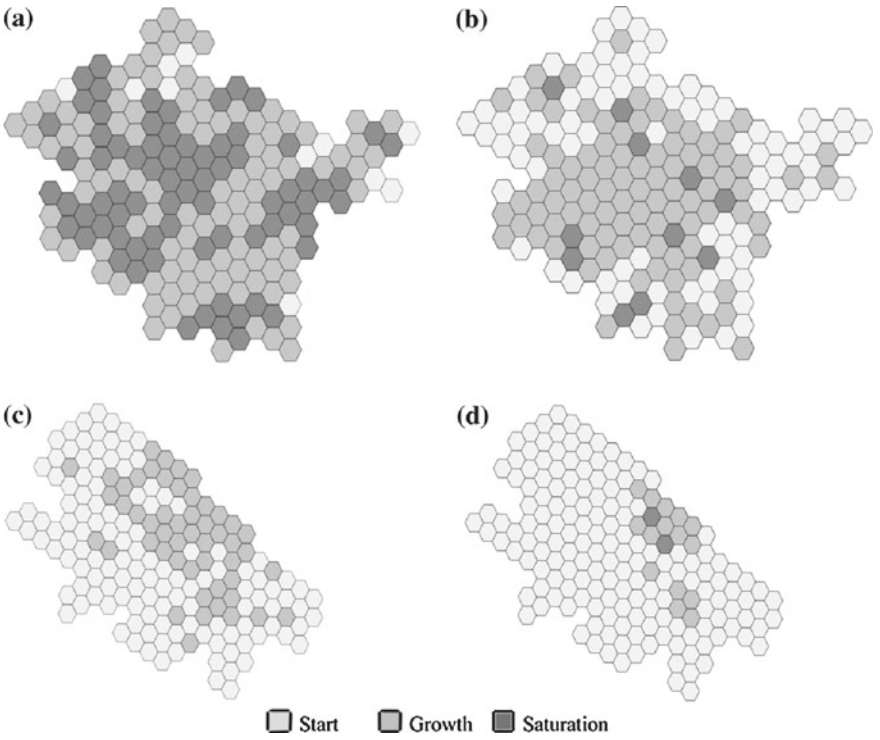


Fig. 6 Comparison between two different progress patterns in Paris and Buenos Aires based on calculated activity stages for the year 2012 using the feature classes *building* and *house number*. **a** Paris—*building*, **b** Paris—*house number*, **c** Buenos Aires—*building*, **d** Buenos Aires—*house number*

rural regions, sparsely populated regions or other feature classes than streets and buildings would be proper candidates. Finally, externally estimated complete regions may be considered as candidates for deriving and validating additional inference rules for completeness estimations. These rules can be used for automated completeness assessments in the future.

Acknowledgments This work was partly funded by the Austrian Federal Ministry for Transport, Innovation and Technology. We thank Pascal Neis for providing the delimitations of the twelve world regions defined in Neis et al. (2013).

References

Arsanjani JJ, Helbich M, Bakillah M, Loos L (2013) The emergence and evolution of OpenStreetMap: a cellular automata approach. Int J Digit Earth 1–30

Barron C, Neis P, Zipf A (2014) A comprehensive framework for intrinsic OpenStreetMap quality analysis. Trans GIS. doi:10.1111/tgis.12073

Corcoran P, Mooney P, Bertolotto M (2013) Analysing the growth of OpenStreetMap networks. Spat Stat 3:21–32
Girres J, Touya G (2010) Quality assessment of the french OpenStreetMap dataset. Trans GIS 14(4):435–459
Goodchild MF (2007) Citizens as sensors: the world of volunteered geography. GeoJournal 69: 211–221
Goodchild MF, Li L (2012) Assuring the quality of volunteered geographic information. Spat Stat 1:110–120
Hagenauer J, Helbich M (2012) Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks. Int J Geogr Inf Sci 26(6):963–982
Haklay M, Weber P (2008) OpenStreetMap: user-generated street maps. IEEE Pervasive Comput 7(4):12–18
Haklay M (2010) How good is volunteered geographical information? a comparative study of OpenStreetMap and ordnance survey datasets. Environ Plan B, Plan Des 37(4):682–703
Haklay M, Basiouka S, Antoniou V, Ather A (2010) How many volunteers does it take to map an area well? the validity of linus' law to volunteered geographic information. Cartographic J 47(4):315–322
Hecht R, Kunze C, Hahmann S (2013) Measuring completeness of building footprints in OpenStreetMap over space and time. ISPRS Int J Geo-Inf 2(4):1066–1091
ISO (2011) Geographic information—data quality (ISO/DIS 19157:2011)
Mondzech J, Sester M (2011) Quality analysis of OpenStreetMap data based on application needs. Cartographica 46(2):115–126
Mooney P, Corcoran P, Winstanley AC (2010) Towards quality metrics for OpenStreetMap. In: 18th ACM SIGSPATIAL international conference on advances in geographic information systems
Neis P, Zielstra D, Zipf A (2012) The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. Future Internet 4(1):1–21
Neis P, Zielstra D, Zipf A (2013) Comparison of volunteered geographic information data contributions and community development for selected world regions. Future Internet 5(2):282–300
OSM Wiki (2013a) Map features. http://wiki.openstreetmap.org/wiki/Map_Features. Accessed 14 Nov 2013
OSM Wiki (2013b) Vienna OSM coverage. http://wiki.openstreetmap.org/wiki/Vienna_OSM_Coverage. Accessed 03 Dec 2013
Rehrl K, Gröchenig S, Hochmair H, Leitinger S, Steinmann R, Wagner A (2012) A conceptual model for analyzing contribution patterns in the context of VGI. In: LBS 2012–9th symposium on location based services. Springer, Berlin
Steinmann R, Brunauer R, Gröchenig S, Rehrl K (2013a) Wie aktiv sind freiwillige Mapper? In: Angewandte Geoinformatik 2013. Beiträge zum 25. AGIT-Symbosium Salzburg, pp 173–182
Steinmann R, Gröchenig S, Rehrl K, Brunauer R (2013b) Contribution profiles of voluntary mappers in OpenStreetMap. In: Online proceedings of the international workshop on action and interaction in volunteered geographic information, 16th AGILE conference
Strano E, Nicosia V, Porta S, Barthélemy M (2012) Elementary processes governing the evolution of road networks. Sci Rep 2:296
Suh B, Convertino G, Chi EH, Pirolli P (2009) The singularity is not near: slowing growth of Wikipedia. In: WikiSym '09 proceedings of the 5th international symposium on Wikis and open collaboration
Zielstra D, Hochmair HH (2011) A comparative study of pedestrian accessibility to transit stations using free and proprietary network data. J Transp Res Board 2117:145–152