# TIME-RELATED QUALITY DIMENSIONS OF URBAN REMOTELY SENSED BIG DATA

Zs. Kugler[1*], Gy. Szabó[1], H. M. Abdulmuttalib[2], C. Batini[4], H. Shen[5], A. Barsi[1], G. Huang[3]

[1] Dept. of Photogrammetry and Geoinformatics, Budapest University of Technology and Economics, Hungary –
(barsi.arpad, kugler.zsofia, szabo.gyorgy)@epito.bme.hu
[2] Dubai Municipality, Dubai, UAE – husseinma@dm.gov.ae
[3] Chinese Academy of Surveying and Mapping, Beijing, China – huang.guoman@casm.ac.cn
[4] University of Milano-Bicocca, Italy – batini@disco.unimib.it
[5] School of Resource and Environmental Sciences, Wuhan University, China – shenhf@whu.edu.cn

**Commission IV, ICWG III/IVb and WG IV/4**

**KEY WORDS:** data quality, data dimensions, quality metrics, time, big data, crowd source

## ABSTRACT:

Our rapidly changing world requires new sources of image based information. The quickly changing urban areas, the maintenance and management of smart cities cannot only rely on traditional techniques based on remotely sensed data, but also new and progressive techniques must be involved. Among these technologies the volunteer based solutions are getting higher importance, like crowd-sourced image evaluations, mapping by satellite based positioning techniques or even observations done by unskilled people. Location based intelligence has become an everyday practice of our life. It is quite enough to mention the weather forecast and traffic monitoring applications, where everybody can act as an observer and acquired data – despite their heterogeneity in quality – provide great value. Such value intuitively increases when data are of better quality. In the age of visualization, real-time imaging, big data and crowd-sourced spatial data have revolutionary transformed our general applications. Most important factors of location based decisions are the time-related quality parameters of the used data. In this paper several time-related data quality dimensions and terms are defined. The paper analyses the time sensitive data characteristics of image-based crowd-sourced big data, presents quality challenges and perspectives of the users. The data quality analyses focus not only on the dimensions, but are also extended to quality related elements, metrics. The paper discusses the connection of data acquisition and processing techniques, considering even the big data aspects. The paper contains not only theoretical sections, strong practice-oriented examples on detecting quality problems are also covered. Some illustrative examples are the OpenStreetMap (OSM), where the development of urbanization and the increasing process of involving volunteers can be studied. This framework is continuing the previous activities of the Remote Sensing Data Quality Working Group (ICWGIII/IVb) of the ISPRS in the topic focusing on the temporal variety of our urban environment.

## 1. INTRODUCTION

Traditional means of data acquisition is usually carried out by remote sensing (RS) industry, government agencies such as national mapping agencies, surveying industry. Their data acquisition methods are usually well documented, and data quality information is provided together with the data. The new technology of crowd-sourcing has opened a new wide area in spatial data acquisition. In contrast, these methods have usually less documented means of acquisition. They carry more uncertainty in quality measures. The trust of their sources is much lower compared to the above mentioned traditional techniques. Still they carry a vast potential that traditional data sources do not.

The rapid development of urban environment requires the tracking of fast changes in data acquisition. Crowd-sourced remotely sensed data may comply to this request by enabling the mapping of the rapidly changing environment, while traditional surveying techniques in many cases take up too much time to work effectively. Big data is nowadays a rapidly growing area of data processing. It is characterized by the 4V-laws: big data has extreme *Volume* (very much data), *Velocity* (it is captured very quickly), *Variety* (big data has very different types and nature), *Veracity* (data quality varies greatly). In remote sensing and geographic information systems there are a lot of areas where big data and related analysis techniques can be involved, moreover this combination has advantages in comparison to the traditional methods. Land cover and land use mapping is an example of such an area, especially focusing on traffic data acquisition. Google traffic information (GoogleTraffic, 2018) is maybe the most known example, but also transportation networks and the corresponding base maps have been created by crowd-sourced big data collection and analysis techniques in the OpenStreetMap project. (OpenStreetMap, 2018)

## 2. URBAN REMOTELY SENSED DATA

There is a strong interrelationship between quality measures and the types of data sources. Data sources selection and collection have strong influence on the remote sensing data quality (RSDQ) dimensions to be used in the process. In order to contribute to this issue, in this paper we focus on big data sources in the domain of remote sensing. The area of crowd-sourcing internet technology has opened new perspectives to remotely sensed information collection and processing. Traditional methods of data acquisition have been extended with innovative means based on non-expert spatial data gathering.

[*] Corresponding author

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLII-4-315-2018

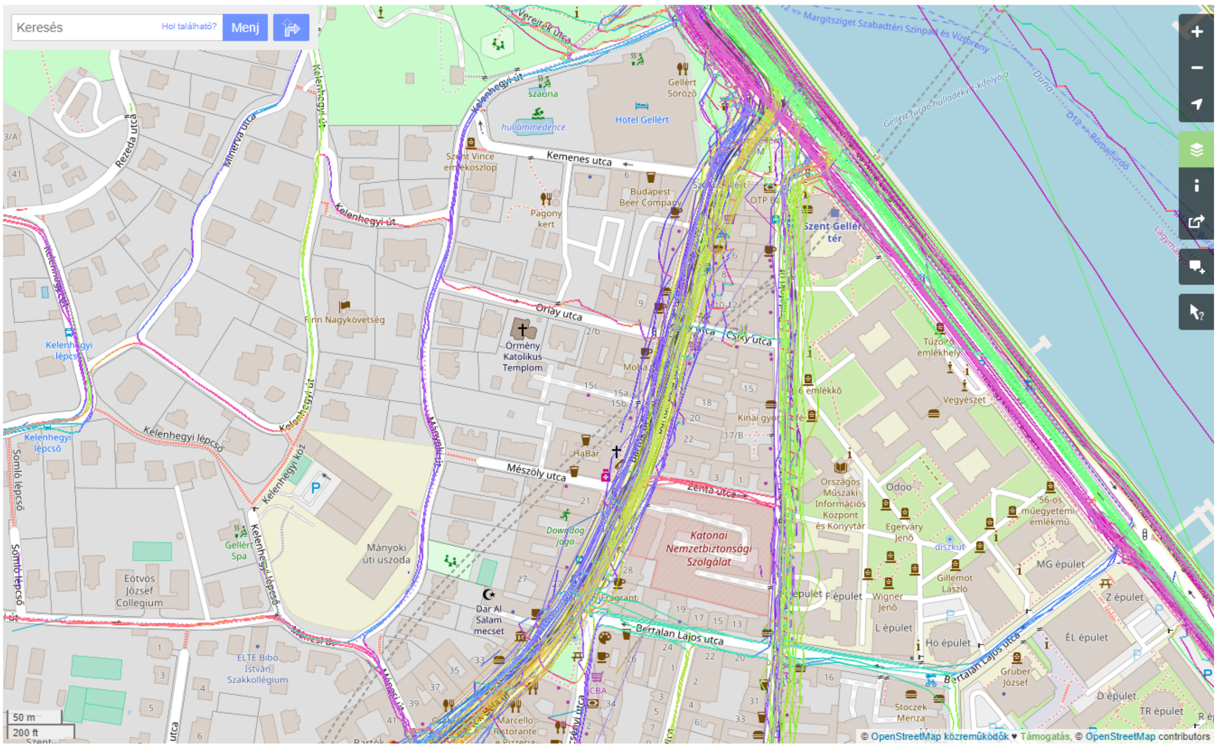

Figure 1. OpenStreetMap detail near to the Budapest University of Technology and Economics campus, Hungary. The color trajectories are from GPS, all the other parts (parcels, buildings, vegetation areas) from imagery evaluation

In the OpenStreetMap initiative, crowd-sourcing of spatial databases is based on aerial and satellite based optical data to obtain geographic data. Source of the spatial information is similar to the one of institutional data production chain.

However, the method adopted to derive geographic information from remotely sensed data is taking a different direction compared to traditional means. Deriving spatial data is on a voluntary basis and performed often by non-expert analysts with lower control on data quality during the production phase. Scarce information is available on quality measures such as trust of sources or consistency of data. Still its great value lies in fast renewable, open accessible nature of the data.

Fig. 1 proves that remote sensing data processing and crowd-sourcing can be integrated smoothly. The area illustrated lies in Budapest, near to the campus of Budapest University of Technology and Economics **(**BME). The road axes are acquired with GPS and similar satellite based measurements, but the area around (buildings, places, etc.) has been obtained using remotely sensed imagery.

Closed courtyards of 4-5 floor buildings along the main roads cannot be mapped by other techniques, but by aerial or satellite image interpretation.

Another promising crowd-sourced big data capturing process is exploited by the assisted and autonomous vehicle technologies (Toth et al, 2018). There is already a pioneer approach – called self-healing mapping technology – to collect environmental data by these special vehicles. Captured data set is transferred into the cloud, and after sophisticated processing they are fed back into the map database (Here, 2018).

## 3. TIME-RELATED QUALITY DIMENSIONS

### 3.1. Terminology

Prior to the discussion of the quality dimensions, some relevant definitions must be given.

- **Time**: „The indefinite continued progress of existence and events in the past, present, and future regarded as a whole." (https://en.oxforddictionaries.com/definition/time) Time is a fundamental scalar quantity, what a clock reads. (Considine, 1985) and „one-dimensional subspace of space-time, which is locally orthogonal to space" (IEC, 2011)
- **Time scale**: „system of ordered marks which can be attributed to instants on the time axis, one instant being chosen as the origin" (IEC, 2011)
- **Time axis**: „mathematical representation of the succession in time of instantaneous events along a unique axis" (IEC, 2011)
- **Event**: „something that happens in subspace time of space-time" (IEC, 2011)
- **Instant**: „point on the time axis" (IEC, 2011)
- **Time interval**: „part of the time axis limited by two instants" (IEC, 2011)
- **Duration**: „range of a time interval", „a non-negative quantity" The units are minute (1 min = 60 s), hour (1 h = 60 min = 3 600 s), and day (1 d = 24 h = 86 400 s). (IEC, 2011)
- **Date**: „mark attributed to an instant by means of a specified time scale" (IEC, 2011)
- (Temporal) **Frequency**: „the number of repetitions of a periodic process in a unit of time" (Merriam-Webster, 2018)

- **Availability**: „The ratio of the total time a functional unit is capable of being used during a given interval to the length of the interval.” (Wikipedia)
- **Second**: SI base unit of time, 1/86,400 of a day. The SI-definition is: „the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the caesium 133 atom.” (SI Brochure)
- **Hertz**: „one cycle per second”, „The dimension of the unit hertz is 1/time. Expressed in base SI units it is 1/second (1/s).” (Wikipedia)
- **Clock** is an instrument used to measure, keep, and indicate time. (Clock, 2018)

### 3.2. Time-related quality dimensions and metrics

Remote sensing data quality has been described for traditional data sources in our previous papers (Albrecht, 2018, Barsi, 2018) As to information technology (Batini, 2016) is a good summary of dimensions and related metrics for different types of data, such as relational tables, maps, images, linked open data, loosely structured texts and laws. Due to the Variety-law, the Big Data domain has a different nature compared to traditional remote sensing data sources and related quality dimensions. Due to high frequency of data collection and faster acquisition methods when compared to traditional means, time-related quality dimensions have an increased emphasis. E.g. while the resolution dimensions are highly relevant, accuracy has much lower significance in case of RS big data sources. The most important dimensions and their metrics are listed in the following table grouped by main data quality clusters after (Batini, 2016):

| **Dimension name** | **Short definition** | **Dimension metric** |
|---|---|---|
| | Resolution cluster | |
| Temporal resolution | the precision of a measurement with respect to time;<br>the sampling frequency | sec, hour, day, week, year |
| | Accuracy cluster | |
| Temporal precision | a measure of the detail in which the date is expressed | sec, hour, day, week, year |
| Currency | how promptly data are updated with respect to changes;<br>validity of data with respect to time | sec, hour, day, week, year |
| Temporal validity | Time interval during which data remain accurate and bring value when used for a purpose | sec, hour, day, week, year |
| Volatility | characterizes the frequency with which data vary in time | sec, hour, day, week, year |
| | Completeness cluster | |
| Temporal completeness | the extent to which the data represents the real world at a given instant | % of objects represented at the given instant, compared to the objects of the universe |
| | Redundancy cluster | |
| Minimality | every part of the data set is represented only once and none could be eliminated | |
| Temporal redundancy | more data are captured than its minimality along a time axis | |
| | Accessibility cluster | |
| Timeliness | how current the data are for the task at hand (or are too late) | [0,1] dimensionless<br>$\max\left(0{,}1-\frac{currency}{volatility}\right)$ |
| | Consistency cluster | |
| Temporal consistency | the violation in time of semantic rules defined over (a set of) data;<br>correctness of ordered events or sequences | |

Table 1. Time-related data quality dimensions and their metrics

Remote sensing uses two common terms to describe environmental phenomena. The first is **instant** while an event occurs and date and time is assigned (e.g. a landslide in the Alps occurred at 8:00 on 10 June 2013). An instant has accuracy, resolution and precision as quality dimensions. The second is **duration**, i.e. how long does an observed event last (e.g. floods occurred in Budapest on the Danube occurred from 10 June to 18 June 2013). Duration can be characterized by accuracy and size (which means the length of the phenomena). Beyond these two terms, repetitive observations are described by frequency, having accuracy and stability as quality dimensions. The latter means how stable/constant is the measurement rate during the data acquisition procedure. Such repeating observations must also be featured by a measure how representative it is, i.e. whether the captured data are suitable to describe the monitored event. Such frequency type measure results in the comparison to the Nyquist rate (which is a sampling rate resulting alias-free signals) (Nyquist-rate, 2018).

## 4. CASE STUDIES

### 4.1. Example 1: Traffic information

The OpenStreetMap (OSM) is one of the most commonly known crowd-sourced databases. The project has started in August 2004. This collaborative mapping aimed to involve skilled professionals and non-professionals to create a map database covering the built-up and rural environment with the main features like waters, roads, land use, buildings and many others. At the starting of the project low-budget portable satellite navigation systems have been used; since 2006, thanks to Yahoo’s collaboration in the project, even aerial photographs can also be interpreted for data capturing. In 2010 Bing also allowed to use satellite imagery in map making. In 2012 Google Maps led several prominent websites, like Foursquare or Craigslist to switch from their service to OpenStreetMap. (OpenStreetMap, 2018) The basic statistics about OSM is the following as of 12 June 2018:

| | |
|---|---|
| Number of users | 4.717.453 |
| Number of uploaded GPS points | 6.242.407.083 |
| Number of nodes | 4.611.782.195 |
| Number of ways | 508.857.430 |
| Number of relations | 5.924.988 |

Table 2: Statistics of the OpenStreetMap

The growth of this giant database shows an interesting path. Raw data are available in tagged XML format or after some conversions and layer creation in shape format. Take the example of the Hungarian OSM development in the last decade. Thanks to the German provider Geofabrik, the database has been yearly downloaded since 2010. The vector data in shape format representation has 18 layers and 91 files containing all the relevant themes such as waters, railways, roads, POIs etc.

Fig. 2 shows the sizes of the shape files together with the yearly increase rates, the amounts are in MB. The dotted lines show the linear trends in growth. While the total information storage was 17.7 MB in 2010, the size of the database has increased by about 44 times, being 785.6 MB this year.

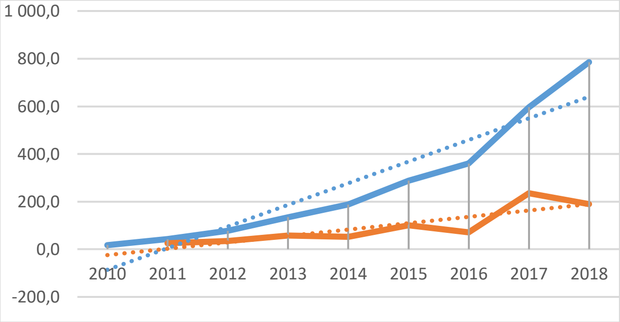

Figure 2. The total size of shape file and its yearly increase rates from 2010 to 2018. Blue represents the database size, orange the yearly increase

The most rapid growth in the map database content can be experienced at the roads layer. Fig. 3 shows the growth of the amount of points and polylines.

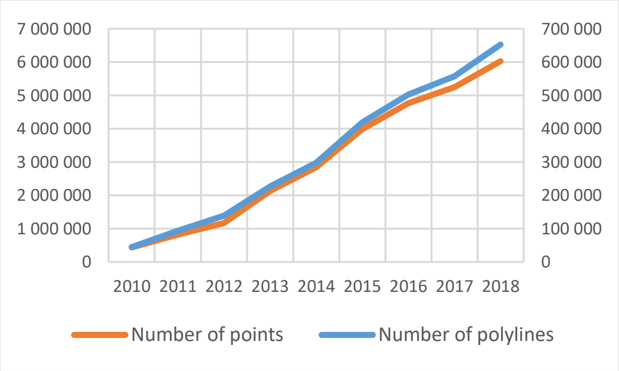

Figure 3. Growth of the road layer

The best visualization of the database evolution is the development of the map visualization of the data. Fig. 4 shows the road network density near to Budapest in years 2010, 2015 and 2018 respectively. Notice that first main roads have been mapped, then lower category roads are digitized.

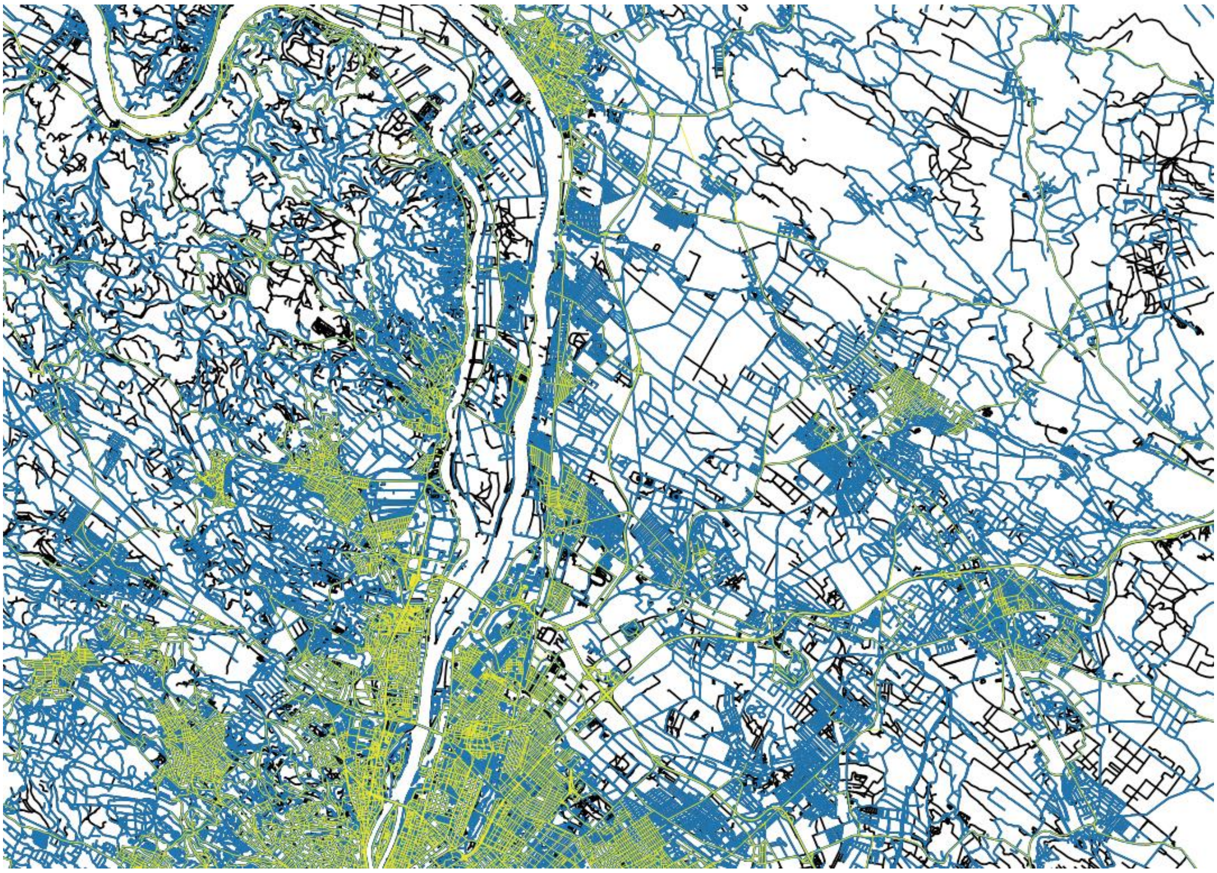

Figure 4. Evolution of the road network density near Budapest in years 2010 (yellow), 2015 (blue) and 2018 (black)

https://doi.org/10.5194/isprs-archives-XLII-4-315-2018 | © Authors 2018. CC BY 4.0 License.

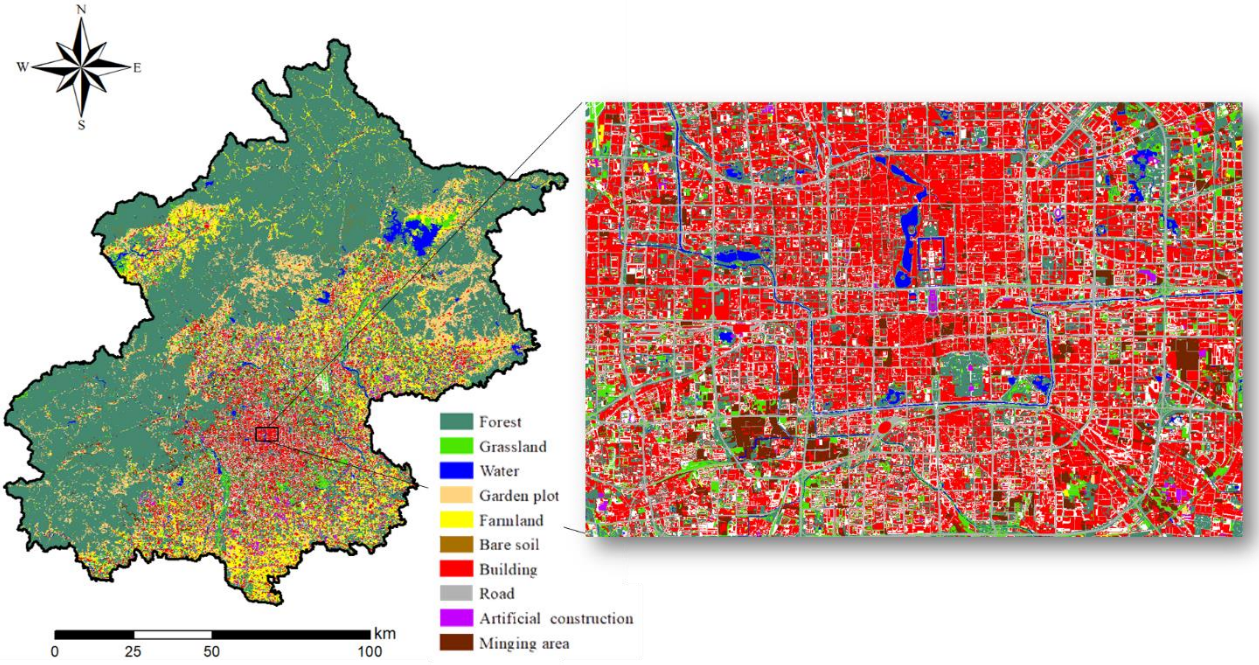

Figure 5. Land cover map of Beijing in 2015 (10 first-level classes)

### 4.2. Example 2: Land cover and land use

In 2011 China launched the Geographical Conditions Monitoring (GCM) project, which aims to reflect the spatial distribution and changing of natural and built-up elements of the environment. It comprehensively utilizes modern mapping and geographic information technology to dynamically and quantitatively monitor land use/land cover. The GCM project divides land surface categories into 10 first-level classes, 35 second-level classes, and 135 third-level classes (Fig. 5). The minimum mapping unit is 400 $m^2$. The nationwide land cover/land use database contains a huge amount of data, which is produced based on standard mapping procedure using high resolution remote sensing images. The classification accuracy of GCM project is higher than 95%.

In order to ensure the accuracy of the land cover/land use maps, strict quality control is required during the project implementation process. The process includes

> (1) data preparation control, i.e. data source quality, device configuration, and personnel qualification;
> (2) producing quality control, i.e. quality control of image preprocessing and information extraction;
> (3) quality inspection at two levels, i.e. first level is the inspection by image operation department, and second level is by quality supervision and inspection department.

The list of quality controls includes resolution, date, and mathematical basis of remote sensing images; map projection; data format; attribute table; topology; edge accuracy; classification accuracy, and so on.

## 5. CONCLUSION

This paper has opened new perspectives to remote sensing data quality management. The traditional means of RS data acquisition nowadays is extended by new methods of crowd-sourced big data. Its strength lies in the rapid development of such databases in comparison to traditional spatial data collection. For this reason the emphasis of their quality measures essentially differs from usual RSDQ. Instead of resolution or accuracy the time related dimensions are the most important measures to evaluate quality. Crowd-source remotely sensed big data like OSM is a good example of the usefulness and the fitness for use of this kind of data.

Knowing the weakness of crowd-sourced data we should not misinterpret the strength of traditional RS data acquisition methods. The future perspective is most likely that both RS data collection methods will extend each other and together provide a strong basis for different application of spatial databases.

## REFERENCES

Albrecht, F., Blaschke, T., Lang, S., Abdulmutalib, H.M., Szabó, G., Barsi, Á., Batini, C., Bartsch, A., Kugler, Zs., Tiede, D., Huang, G., 2018. Providing Data Quality Information for Remote Sensing Applications, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLII-3, pp. 15-22.

Barsi, Á., Kugler, Zs, László, I., Szabó, Gy, Abdulmutalib, H.M., 2018. Accuracy Dimensions in Remote Sensing, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLII-3, pp. 61-67.

Batini, C., Blaschke, T., Lang, S., Albrecht, F., Abdulmutalib, H. M., Barsi, Á., Szabó, G., and Kugler, Zs., 2017. Data Quality in Remote Sensing, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLII-2/W7, pp. 447-453, doi.org/10.5194/isprs-archives-XLII-2-W7-447-2017

C. Batini, M. Scannapieco, 2016. Data and Information Quality, Data-Centric Systems and Applications, Springer International Publishing Switzerland, p.500, doi.org/10.1007/978-3-319-24106-7_2
Clock, 2018. https://en.wikipedia.org/wiki/Clock (25 June 2018)

Considine, Douglas M.; Considine, Glenn D., 1985. Process instruments and controls handbook (3 ed.). McGraw-Hill. pp. 18–61.

Frequency definition, 2018. https://www.merriam-webster.com/dictionary/frequency (25 June 2018)

Google Traffic, 2018.
https://en.wikipedia.org/wiki/Google_Traffic (25 June 2018)

Here Technologies, 2018. Self-healing Map,
https://engage.here.com/self-healing-map (25 June 2018)

IEC 60050-113:2011, International Electrotechnical Vocabulary - Part 113: Physics for Electrotechnology, 2011

Li Deren, Ding Lin, Shao Zhenfeng. 2016. Reflections on Issues in National Geographical Conditions Monitoring. *Geomatics and Information Science of Wuhan University*, 41(2), pp. 143-147.

Li Deren, Qi Haigang, Shan Jie, 2012. Discussion on Key Technologies of Geographic National Conditions Monitoring. *Geomatics and Information Science of Wuhan University*, 37(5), pp. 505-512.

Li Guangdong, Fang Chuanglin, Wang Shaojian, et al., 2016. Progress in remote sensing recognition and Spatio-temporal Changes Study of Urban and Rural land use. *Journal of Natural Resources*, 31(4), pp. 703-718.

Li Xia, Li Dan, Liu Xiaoping, 2017. Geographical Simulation and Optimization System (GeoSOS) and Its Application in the Analysis of Geographic National Conditions. *Acta Geodaetica et Cartographica Sinica*, 46(10), pp. 1598-1608., doi.org/10.11947/j.AGCS.2017.20170355.

Nyquist-rate, 2018. https://en.wikipedia.org/wiki/Nyquist_rate (25 June 2018)

OpenStreetMap, 2018.
https://en.wikipedia.org/wiki/OpenStreetMap (25 June 2018)

SI Brochure, 2014. The International System of Units (SI) [8th edition, 2006; updated in 2014]

Toth, C.K., Koppanyi, Z., Lenzano, M.G., 2018. New Source of Geospatial Data: Crowdsensing by Assisted and Autonomous Vehicle Technologies, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, Vol. XLII-4/W8, pp. 211-216.