reviewed paper

Smart Urban Simulation Tools for Planning Decision Support Need Smart Data and Smart Data Gathering Methods

Ernst Gebetsroither-Geringer, Wolfgang Loibl, Mario Köstl, Jan Peters-Anders

(Dr. Ernst Gebetsroither-Geringer, AIT Austrian Institute of Technology, Giefinggasse 2, 1210 Vienna, ernst.gebetsroither@ait.ac.at)
(Dr. Wolfgang Loibl, MSc, AIT Austrian Institute of Technology, Giefinggasse 2, 1210 Vienna, wolfgang.loibl@ait.ac.at)
(Mag. Mario Köstl, AIT Austrian Institute of Technology, Giefinggasse 2, 1210 Vienna, mario.koestl@ait.ac.at)
(Mag. Jan Peters-Anders, AIT Austrian Institute of Technology, Giefinggasse 2, 1210 Vienna, jan.peters-anders@ait.ac.at)

1 ABSTRACT

Urban growth is a challenge for most cities all over the world, especially in less developed countries. This tendency is calling for for smart/innovative instruments to foster sustainable urban development. Decision support for urban planning is required in order to reduce costs and resources to better accommodate new population, willing to move into urban areas. Latin American countries e.g. went from being predominantly rural to predominantly urban within a few decades, leading to high concentrations of urban population. This urban growth is expected to continue leading to severe financial stress for city budgets in order to provide the required infrastructure. AIT - Austrian Institute of Technology has been contracted by the Inter-American Development Bank (IDB) to develop a smart "Urban Infrastructure Development Simulator" (UIDS) – a tool able to performe urban growth simulation and related infrastructure cost estimations, which can be used to support urban planning decisions. In order to enable the cities to make their decisions an Agent-based simulation model has been developed representing the urban growth by estimating dwelling behaviour of the cities' current residents and future residents coming from urban regions outside the city. The urban growth simulation tool is based on input data of different spatial and temporal resolution. Data from Geographical Information Systems (GIS), remote sensing data as well as statistical data are used to simulate scenarios for future development paths. To support the urban planning process such kind of tools need to have great flexibility concerning their data management, e.g. in providing different possibilities to import new (e.g. more accurate) data to calculate new scenarios. Beyond this common need, questions arise like: What happens if the data is not or only partially available and how might a data gathering process be supported by new tools and methods? This paper will introduce different innovative ways how urban planners might be supported to gain new data, which can be used in tools like the UIDS. The developed approaches enable urban planners to easily introduce important tacit knowledge about their city into the simulation tool. Additionally, a method will be depicted how citizens can be enabled to participate in the collection of such data. The paper will further elaborate on challenges the UIDS team encountered and on solutions to overcome these problems using data of different temporal and spatial resolution. The results depicted in this paper are based on experience gathered whithin several urban growth simulation projects performed for different regions in Europe and Latin America.

2 INTRODUCTION

Urban areas can be seen as innovation ecosystems wherein solutions are created or deployed to accelerate the most often aimed transition to a more sustainable, resource efficient urban system. Citizens in this ecosystems can be pro-active catalyzers of innovation, shaping cities as actors of change.

Decision support systems, such as the one presented in this paper, are built to facilitate urban design processes. They may aim at providing the local government with knowledge about citizens' preferences in order to consider and/or include those preferences in the decision-making process for urban development plans. Preferences of, e.g., where to live or move within the city, can be visualized with scenario simulations using Agent-based modelling (ABM).

It is not enough, though, to build smart decission support tools, which are in principle able to facilitate the decisison whithout appropriate data to feed them. Over the last few years our experienc in this context has shown that it is not an easy task to define what kind of data should be used. One important challenge in this kind of simulation is how to gather citizens' preferences which can be used to retrieve the behavioral rules required for ABMs. There are different ways how this can be achieved: One way is extracting information through static data analysis. The downfall of this strategy is that data is often either not available in the required resolution/detail or not available at all and if information is available it might be outdated and therefore not useable. This paper will discuss the benefits of a different, smarter approach of gathering data, i.e. a participatory data gathering procedure.

The improvements in data analyses and data collection methods have been tremendous during the last few decades, nevertheless, especially in the context of analysing past trends, this new and often called smarter data is by no means per definition smarter. The focus of this paper lies in procedures that use remote sensing methods to gather new data. Today, these methods can record data in a spatial resolution of 1 to 1,000 m² cell sizes, but does a higher resolution always produce better information and how can a higher resolution be upscaled, if necessary? These are questions which will –at least briefly- be discussed in the following.

3 SMART URBAN SIMULATION TOOLS

URBANICA, formerly called Urban develoment and infrastructure cost calculator (UIDS), is a decision support tool based on several years of experience in urban growth simulation. It is currently under further development for the Inter-American Development Bank (IADB). The development of URBANICA started in 2014. Since then several different versions (prototypes) have been developed (Gebetsroither-Geringer and Loibl, 2014; Gebetsroither-Geringer and Loibl, 2015). But the question is: What makes a tool like URBANICA smart? In our perspective there are a few main characteristics a smart tool has to consider:

(i) Smart tools need to find a balance between all the features they can/need to provide and the necessary amount of time users need to get results from the tool. This challenge can be tackled by software development in close cooperation with the end user as well as creating different versions of the tool, i.e. to make a simpler version for standard users and an expert version for advanced users.

(ii) Another characteristic is to be flexible in the kind of data which can be fed into the tool or in the formats the results can be exported to. In the case of URBANICA, GeoTiffs, ESRI Shapefiles, KMZ files (Google Earth overlays), images (Portable Network Graphics (PNGs)) and CSV files are the most valuable ones.

(iii) A third characteristic is to be fast it creating results. Experince showed that users do not want to wait too long to see the results of their proposed urban planning decisions. URBANICA e.g. can calculate standard scenarios, simulating 20 years, whithin 2-3 minutes of calculation time. Important in this context is that it is not only a question of absolute calculation time, it is also the perceived impression of the user if nothing obvious is happening and they feel bored.

(iv) A last challenge is to take into account, on the one hand, the user’s wish that every aspect of the simulation can be influenced, meaning, e.g., that -at best- all parameter settings can be changed manually, but on the other hand to have one “perfect” single solution (one proper decision) at the end, which hardly is the reality. If the latter is the case, other tools are often developed as “black boxes” with no insight into the “mechanisms” of the box and if the former is the case then the users often do not know how to decide what to do since the degrees of freedom are too high. Both extremes are not perceived as smart, nevertheless these are challenges model developers can hardly overcome.

The above list of characteristics is just representing the main challenges we have been facing during the development of URBANICA, it does not contain all the needs a smart decision support tool has to fulfil. But what more is needed?

4 SMART DATA

The following section presents our experience while tackling the challenge of finding appropriate data for URBANICA for different regions within Europe and Latin America. So far, the tool has been applied for four different cities and city-regions, with different data availability/credability.

4.1 The Latest Data is not Always the Smartest Data

The origin of the following challenge is that URBANICA calculates its trends on the basis of different land cover layers of past urban developments and uses these trends to create scenarios for the future. The input data for this procedure had been available at a 30m resolution in the past, but recently an example of a 1.5m (a higher, “better”, “smarter”? resolution) as input for 2013 emerged, accompanied by two layers at 30m resolution (for the years 1986 and 2001). These datasets needed to be compared with each other.

On a 30m resolution it makes sense to define classes representing high, medium and low urban development categories[1], but at 30m it can be hardly determined if a pixel really contains manmade structures or only dry or barren soil (or a combination of them). Satellites delivering such resolutions – e.g. LANDSAT[2] – just allow concluding about the spectral properties of the 30m pixel, of course depending on the wavelength range of their sensors – which in itself is -per definition- a mixed pixel of different "real" objects like trees, buildings or roads. Generally, a single, classified land cover pixel alone does not say anything about the real land cover of this pixel, so it can be hardly estimated, which exact spatial composition is responsible for an actual spectral representation. E.g. it is possible that 50% high sealed soil and 50% grassland would lead to the same spectral 30m properties as a 90% loosely built-up area.

Only in a broader context of several pixels one can decide if a particular pixel is a part of an urban area or any other kind of land cover. So pixel based classifications depend strongly on the rules for this classification, and therefore on the experience of the classifier and the actual method he uses. This on the other hand depends, of course, strongly on the type of sensor that has been used to assess the input. This is, amongst others, one of the disadvantages of a pixel based classification and today one rather uses so called feature based classification methods which allow classification schemes on a vector base. By segmenting the survey and combining pixels with similar properties one gets so called image objects, which represent different land cover types. Nevertheless, if one needs data which is comparable with historical ones, pixel based classification still makes sense, but it has to be guaranteed, that one uses comparable sensors and seasons (e.g. before or after an explicit rainy season) so that the resulting classification is really comparable to older ones. Otherwise a particular region could be classified completely different apparently showing enormous land cover changes, although in reality hardly anything has changed.

Contrary to a 30m resolution, it is clear that on a resolution of just 1.5m one gets completely different content. Such classifications cannot generally be compared to 30m resolution data. At 1.5m it does not make sense to speak about high or low urban intensity, because this pixel representation only allows for a statement whether a soil pixel has a high or low degree of sealing. It is also not possible to decide whether a special pixel is part of a forested area or of open grassland. The only statement possible is that the pixel has a high or low vegetation index, again depending on the spectral properties of the used sensor. At such a high level of detail one should perform an object based image analysis (OBIA) rather than a pixel based analysis (Blaschke, 2010)[3]. Nevertheless, such high (spatial) resolution data sets can definitely present a surplus value, but only as additional data sets allowing to identify interesting regions and to discover why -on a lower resolution- a special land cover class has been identified. This problem is quite severe if these 1.5m (high resolution) classes use the same land cover categories as the 30m ones and should be used to compare different layers to calculate changes.

Of course, one can always try to resample such a classification up to 30m. There are indeed in most GIS platforms (tools like ArcGIS[4] or QGIS[5]) default RESAMPLE operations to do this. For discrete data, such as a land cover datasets, there are two common options:

(i) The nearest (neighbour) method does not change[6] the values of the input layers. It more or less uses the value of the originally central pixel within the new lower resolution cell (i.e. in our example the 30m).

(ii) The majority method determines the new value of the cell based on the most popular values within the filter window and tends to result in a smoother representation than the nearest (neighbour) method. Both have their pros and cons, but in many cases both of them just create new problems. The following figures show why.

[1] degree of soil sealing

[2] https://en.wikipedia.org/wiki/Landsat_program, checked 17.4.2016

[3] see e.g. http://gisgeography.com/image-classification-techniques-remote-sensing/, checked 17.4.2016

[4] https://www.arcgis.com/, checked 17.4.2016

[5] http://www.qgis.org/de/site/, checked 17.4.2016

[6] "A technique for resampling raster data in which the value of each cell in an output raster is calculated using the value of the nearest cell in an input raster. Nearest neighbour assignment does not change any of the values of cells from the input layer; for this reason it is often used to resample categorical or integer data (for example, land use, soil, or forest type), or radiometric values, such as those from remotely sensed images."
http://support.esri.com/en/knowledgebase/GISDictionary/term/nearest%20neighbor%20resampling

Figure 1 shows the original 1.5m input and the result of the nearest option. For better identification of the problems that can occur with this option, the ArcGIS base imagery and a 300m reference raster have been used, as well as three blue marked 30m cells, which have been used to analyse/depict the problems.

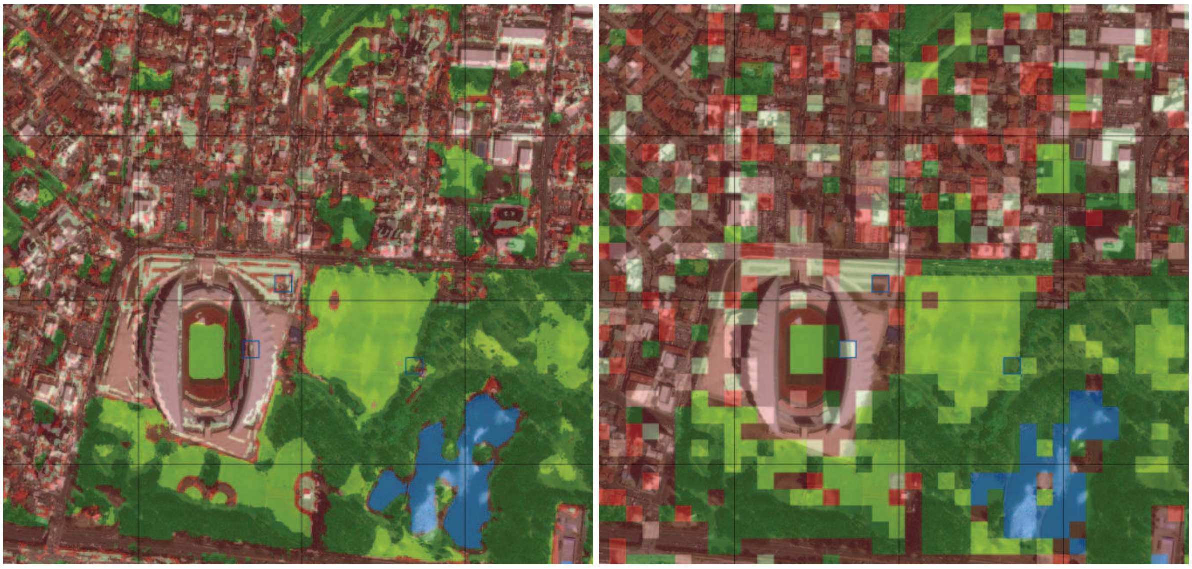

Fig. 1: Exemplary detail of a typical urban region represented by the original 1.5m land cover classification of 2013 (left) and the result of the nearest neighbour resampling to 30m (right)

Regardless of the quality of the 1.5m input the three blue marked cells (see also the white ellipses and black arrows) show very well the possible unexpected resample results using the nearest option. E.g. at the east stand of the stadium we encounter the following: After resampling, a 30m bare land cell occurs, which has not been expected when looking at the input. The reason for this is that the central 1.5m pixel within the new 30m representation is of this type. Although almost all other 1.5m pixels are of the category high density urban, the 30m cell becomes bare land because of the used resampling option. Similarly, in the park area above the legend of the map, a forested land pixel occurs, despite the majority of 1.5m grassland pixels.

With the majority option (method) shown on the left hand side in figure 2, no substantial improvement is achieved. Of course this method seems to generate a more realistic pixel representation – and that is true for this particular urban region – but in general it also intensifies the dominant class high urban density (compare table 1 further below).

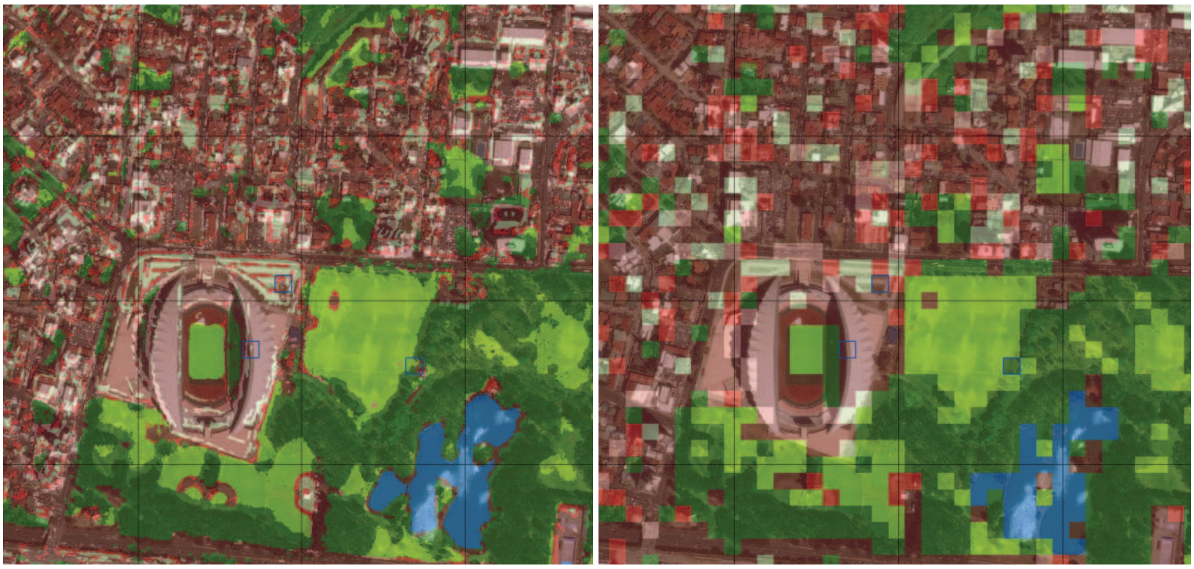

Fig. 2: Exemplary detail of a typical urban region represented by the original 1.5m land cover classification of 2013 (left) and the result of the majority resampling to 30m (right)

If the 1.5m classification uses the same land cover classes – especially for the urban density (high, medium and low) – the resampled cell will never represent a mixture of these classes. E.g. in the case of the above mentioned possible 50% high density urban and 50% grassland distribution, a 30m pixel could only be high density urban or only grassland pixel depending on the actual arrangement of the original 1.5m pixels and the resample option, but never a mixture of these classes - e.g. medium or low density urban as expected with an originally 30m classification.

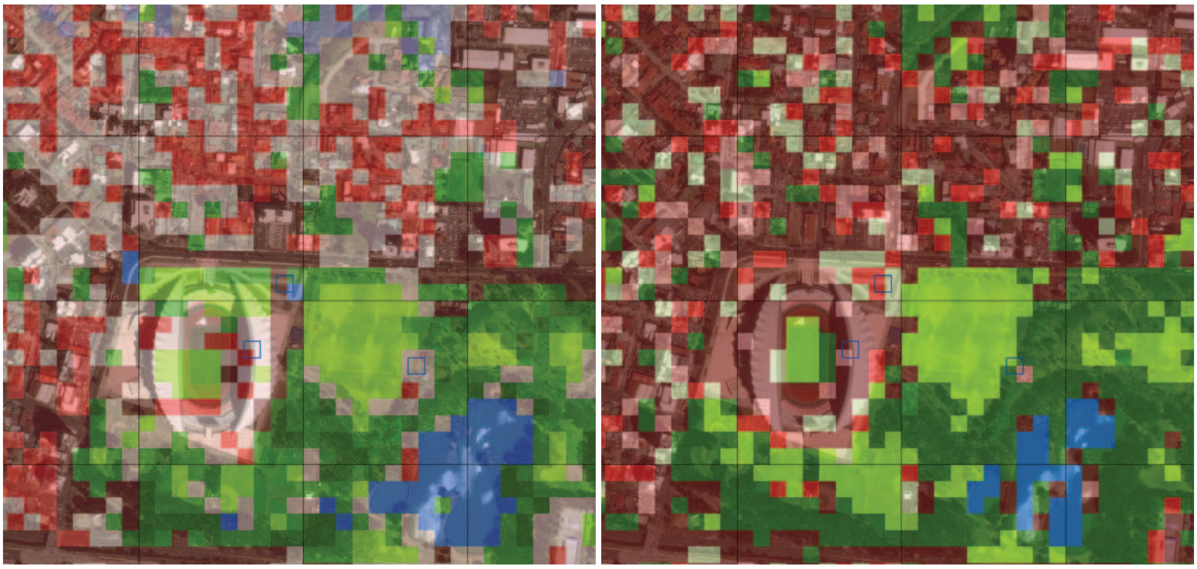

Fig. 3: Re-projected land cover 2001 (left) and resampled land cover 2013 after defining of a common processing extent

Additionally, another problem occurred that the two older land cover layers (1986 and 2001) were created using an entirely different sensor type. Since both, the original 30m data –LANDSAT –as well as the input for the land cover 2013 were not available and also no information about the classification rules and methods, it can only be speculated about the quality of the respective classifications.

We think that all of the above shows quite well that different sensors, resolutions, processing extents and projections for land cover layers should be avoided as far as possible. Otherwise these layers are not really comparable and no conclusions about accurate actual land cover changes can be made.

Thus, the 2013 land cover dataset thematically differs extremely from the two other datasets, although the used classes suggest that this would not be the case. The following table indicates this once again very clearly: While the growth of the total amount of all urban areas in the study area seems to be plausible, the distribution of the particular density classes is very unrealistic.

	study area 1986		**study area 1986**		**study area 1986**	
	hectare	% of urban	hectare	% of urban	hectare	% of urban
1 - high urban density	3558.8	16.8%	4406.0	18.5%	20379.4	78.8%
2 - medium urban density	5890.0	27.8%	6969.3	29.2%	3916.3	15.2%
3 - low urban density	11747.3	55.4%	12491.0	52.3%	1548.9	6.0%
	21195.3	**100%**	**23866.4**	**100%**	**25844.6**	**100%**

Table 1: Comparison of the amount of the three urban density classes of the original land cover layers within the study area

Looking at the distribution of the three urban density classes of the years 1986 and 2001, the order of the several classes is still comparable and the percentage increase of both denser classes at the expense of the third class is very plausible. However, the classification of 2013 draws a different picture: Now, not just about 20% are of high density urban, but almost 80%. The class low density urban on the other hand, which – in both cases – previously accounted for more than 50%, hardly occurs. In our view it is very unlikely that such a compression corresponds with reality. Rather, this comparison shows once more the fundamental incommensurability of the three classifications. As already mentioned above, using the majority resample option this apparent growth (or better: densification) would even be increased.

The following figure 4 shows the main problems once more depicted in a map. The upper panel shows the 2001 representation of the three urban classes, while the middle panel shows the result of the nearest option for the urban classes of 2013. Comparing these two one can easily discern the difference of the content of both classifications. The dark red high density urban pixels predominate in 2013 exorbitantly. For URBANICA this would have the fatal effect that from 2013 onwards just very few pixels would allow further densification, thus leading to an extreme overestimation of the need for new undeveloped areas and therefore unrealistic scenario results. The lower panel shows the result of an alternatively generated 30m land cover layer, developed to solve this problem. This will be explained further below, but in short, the development of this dataset was essentially based on a GIS operation called AGGREGATE (cp. RESAMPLE).

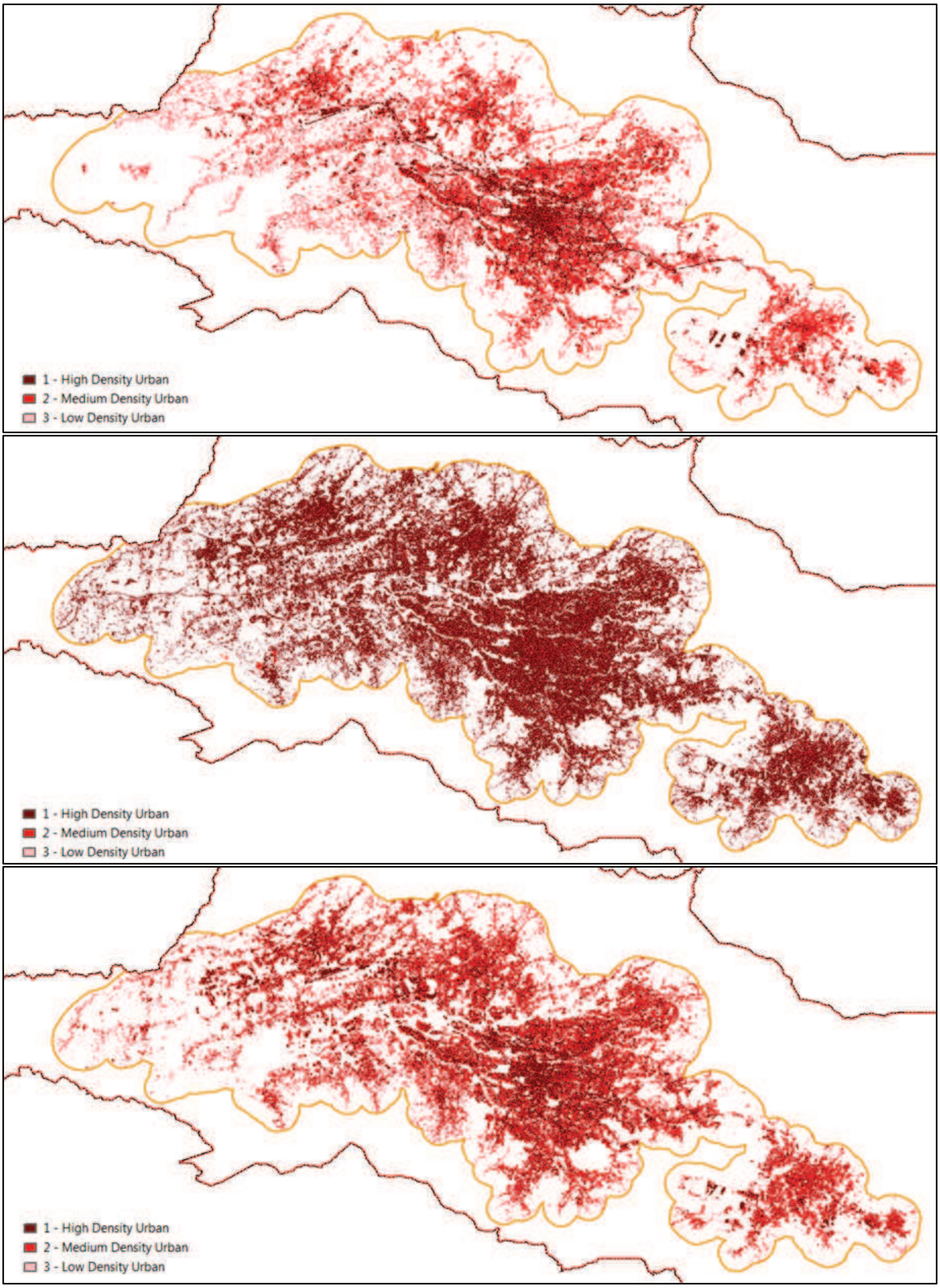

Fig. 4: Comparison of 2001 urban mask (above), the 30m RESAMPLED mask 2013 (middle) and the generated Layer using AGGREGATE for 2013 (below)

The middle panel of the figure above shows that the original situation was more than unsatisfactory. Nevertheless, in order to get a reasonable spatial distribution as input for the tool, we tried to create a more realistic and comparable land cover layer for 2013. As mentioned above, this procedure uses the GIS function AGGREGATE to create a 30m land cover. For this paper it would lead too far to describe this generation process in detail, but substantially the method uses single binary representations of each class. The function generates a reduced-resolution version of a raster (30m instead of 1.5m) by using a factor which is used to multiply the cell size of the input raster to obtain the desired resolution for the output raster. In our case this factor was 20, because the output cell size of 30m is 20 times larger than that of the input raster. We used the SUM option leading to a 30m raster containing the sum of original pixels within. This had to be done for each class separately. At the end we could examine the dominant class simply by division by the factor 400 (= 20 x 20 pixel). Using intelligent rules one can also generate comparable 30m mixture representations of a new cell. We used e.g. up to 33% of "sealed" pixels for low, 33 to 66% for medium and more than 66% for high urban density. Comparing the results of this process visually with the one of the resample method shows the substantial improvement. Now, 29% of the pixels were classified as high density urban, 44% as medium and at least 28% as low. Of course, this is still not a perfect solution: The high proportion of medium density urban areas still seems to be unrealistic, which calls for a further improvement of this approach and consolidates –once again- our warnings to use different data sources at all.

We think this section shows clearly that a certain operating expense is important to get reasonable, appropriate data inputs. Different data coming from different sources may lead to more effort in the end. In the worst case, unrealistic data input might not be detected at all leading to wrong end results. Thus, a higher resolved ("better") data set –using a more recent technology– does not always mean that this data is smarter.

4.2 A Smart Data Gathering Process

For UIDS a new approach was developed to gather data for the Agent-based simulation due to the lack of available data and, henceforth, unsatisfactory results from a common, statistical approach in a project carried out in the City of Ruse, Bulgaria. As this is already described in Gebetsroither-Geringer and Loibl, 2014 and Gebetsroither-Geringer and Loibl, 2016 we want to present here only a summary and a discussion what this data gathering process makes it smarter than others.

The first reason is that city administrations and urban planners are more and more interested in increasing their knowledge about the current preferences of their citizens, which can be hardly derived from data of the past. Processes that can be included in e-governance and e-government[7] were considered as becoming increasingly relevant. The ongoing development of mobile applications supporting this data gathering process increases the amount of available data, but can still be improved, mainly regarding the usability and appropriateness of the gathered data for modelling of urban development. In our approach we used an online questionnaire asking the citizens very few questions. We asked e.g. which areas of the city they:

(a) like most,

(b) could imagine to move to,

(c) do not want to live in at all.

Further we asked for permission to use this information as data input for a simulation to derive attractiveness maps of their city.

The calculation used to derive the attractiveness describes the citizens' attraction to target areas, defined as, e.g., urban raster cells or districts:

$CA_i = f(\sum posPr_i, \sum intPr_i, \sum negPr_i)$ Equation 1

with:

$negPr_i$ = negative preference at target area i

$posPr_i$ = positive preference at target area i

$intPr_i$ = intermediate preference at target area i

CA_i = Citizens' attraction to target area i

[7] eParticipation, 2016

The probability Pi for a target area i to be chosen by a citizen is normalized to 1 for areas of highest attractiveness (i.e., areas where citizens would most probably move to):

$$P_i = CA_i / MAX(CA_1;CA_2;\ldots CA_n) \qquad \text{Equation 2}$$

The derived attractiveness maps were published, e.g., using a Web Map Service (WMS), offering an added value to the citizens who could receive feedback through these maps. Keeping the derived attractiveness maps up to date requires very low effort: E.g., every 1 to 5 years, the same questionnaire could be used and the development since the previous investigation could be visualized. These further advantages make the approach smarter. Details on the approach, the implementation and a comparison to a more commonly used statistical approach can be found in Gebetsroither-Geringer and Loibl, 2016.

5 CONCLUSION AND OUTLOOK

This paper briefly discussed that smart tools (beyond the challenges regarding user-friendliness and the demand for high calculation speeds and credibility) need smart data as input. The example of high resolution remote sensing data is only one example, out of several, wherein supposed data improvements may lead to pitfalls. Thus, in the end, it is not easy to determine what smart data is and this question will always have to be answered on a case by case basis in the context of the data requirements of a software/use case. New data gathering processes are promising and the presented very simple approach will most probably be further extended as e.g. research projects like smarticipate are working on data-rich citizen dialogue systems, transforming public data into new intelligence The project aims to integrate bottom-up processes in the realm of city planning, using the full potential of citizens by sharing ideas in the co-production of decision making. Such kind of projects will open a wide range of new smart data resources, which can and should be used for urban decision support systems like URBANICA.

6 REFERENCES

Berland, M., Rand, W.: Participatory simulation as a tool for agent- based simulation. Matthew Berland Dept. of Computer Sciences/ICES, Univ. of Texas at Austin, USA William Rand Department of Computer Science, Univ. of Maryland, USA. http://www.berland.org/files/berland-icaart09.pdf (2009)

Blaschke T.:. Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing 65, pp.2-16 (2010)

eParticipation: Digital Agenda for Europe https://ec.europa.eu/digital-agenda/en/eparticipation (2017). Accessed 17 April 2016

Gebetsroither-Geringer, Ernst, and Wolfgang Loibl. „Urban Development Simulator: How Can Participatory Data Gathering Support Modeling of Complex Urban Systems“. In Understanding Complex Urban Systems, herausgegeben von Christian Walloth, Ernst Gebetsroither-Geringer, Funda Atun, und Liss C. Werner, 33–47. Understanding Complex Systems. Springer International Publishing, 2016. http://link.springer.com/chapter/10.1007/978-3-319-30178-5_3.

Gebetsroither-Geringer, Ernst. „Multimethod modeling and simulation supporting urban planning decisions“. In Understanding Complex Urban Systems: Multidisciplinary Approaches to Modeling, 13–27. Springer International Publishing, 2014. http://link.springer.com/chapter/10.1007/978-3-319-02996-2_2.

Gebetsroither-Geringer, Ernst, and Wolfgang Loibl. „Urban Development and Infrastructure Cost Modelling for Managing Urban Growth in Latin American Cities“. Last check: 17. April 2016. http://www.corp.at/archive/CORP2015_120.pdf.

Gebetsroither-Geringer, Ernst, und Wolfgang Loibl. „Urban Development Simulator: An Interactive Decision Support Tool for Urban Planners Enabling Citizen’s Participation“. Last check: 17. April 2016. http://geomultimedia.org/archive/CORP2014_149.pdf