UNIVERSITÄT HEIDELBERG
ZUKUNFT SEIT 1386

Enriching crowdsourced land use information by data mining techniques

Jamal Jokar Arsanjani

GIScience research Group. Heidelberg University, DE

GIScience

RICH-VGI workshop,
18th AGILE Conference on Geographic Information Science
Lisboa, Portugal. Tue, Jun, 09, 2015

UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

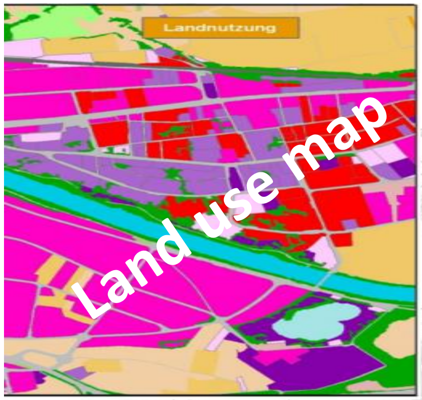
Status of contributed land-use features in OpenStreetMap

Traditional way of land use/cover mapping

Land Surveying (in-field)

Remote sensing

+ Expert knowledge
(in-field, questionnaires, interviews)



Costly

Time-

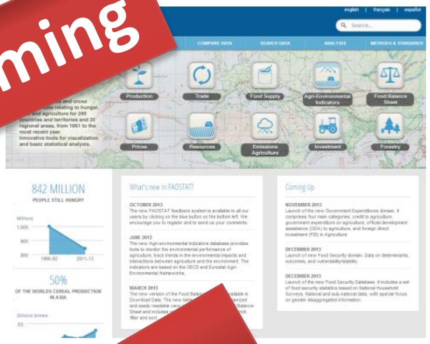
Human-independent

Global/Regional/local datasets

EU: Urban Atlas

GLC2000

Agro-MAPS



AT

Still not accurate enough

Time and cost consuming

Considerable disagreement between them

,....

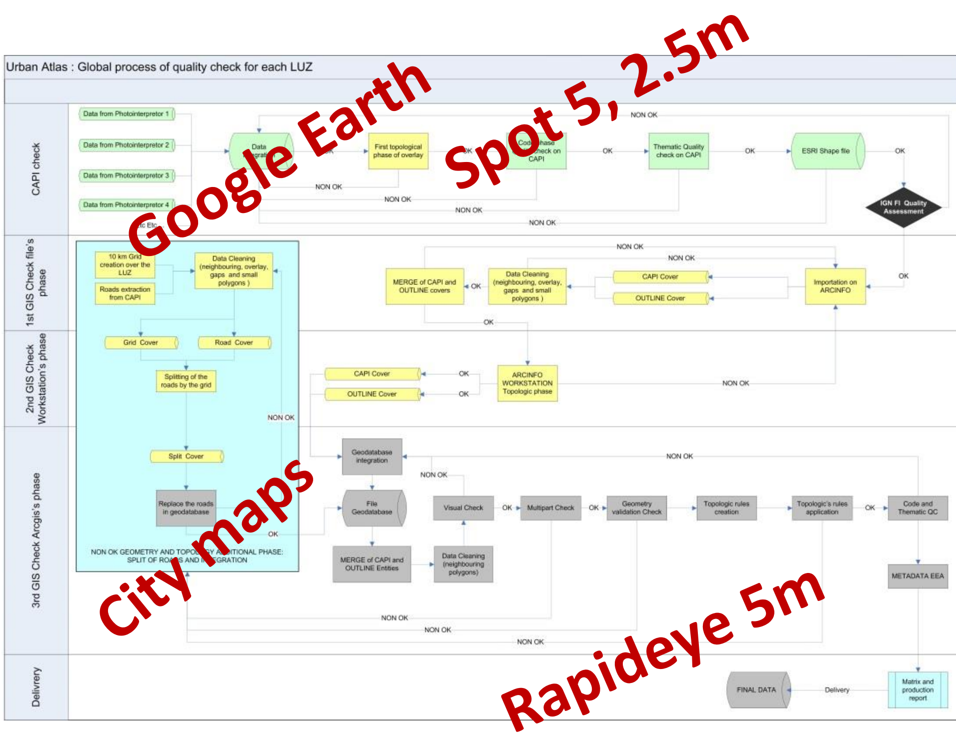
UNEP-WCMC

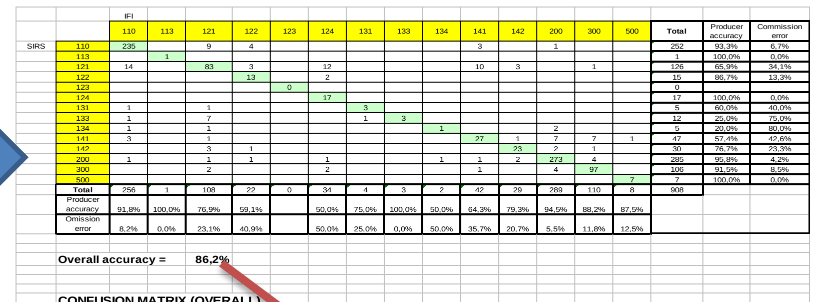
Glo pping

FAO

IFPRI

Process of Urban Atlas data preparation





		IFI																
		110	113	121	122	123	124	131	133	134	141	142	200	300	500	**Total**	Producer accuracy	Commission error
SIRS	110	235		9	4						3		1			252	93,3%	6,7%
	113		1													1	100,0%	0,0%
	121	14		83	3		12				10	3		1		126	65,9%	34,1%
	122				13		2									15	86,7%	13,3%
	123					0										0		
	124						17									17	100,0%	0,0%
	131	1		1				3								5	60,0%	40,0%
	133	1		7				1	3							12	25,0%	75,0%
	134	1		1						1			2			5	20,0%	80,0%
	141	3		1							27	1	7	7	1	47	57,4%	42,6%
	142			3	1							23	2	1		30	76,7%	23,3%
	200	1		1	1		1			1	1	2	273	4		285	95,8%	4,2%
	300			2			2				1		4	97		106	91,5%	8,5%
	500														7	7	100,0%	0,0%
	Total	256	1	108	22	0	34	4	3	2	42	29	289	110	8	908		
	Producer accuracy	91,8%	100,0%	76,9%	59,1%		50,0%	75,0%	100,0%	50,0%	64,3%	79,3%	94,5%	88,2%	87,5%			
	Omission error	8,2%	0,0%	23,1%	40,9%		50,0%	25,0%	0,0%	50,0%	35,7%	20,7%	5,5%	11,8%	12,5%			

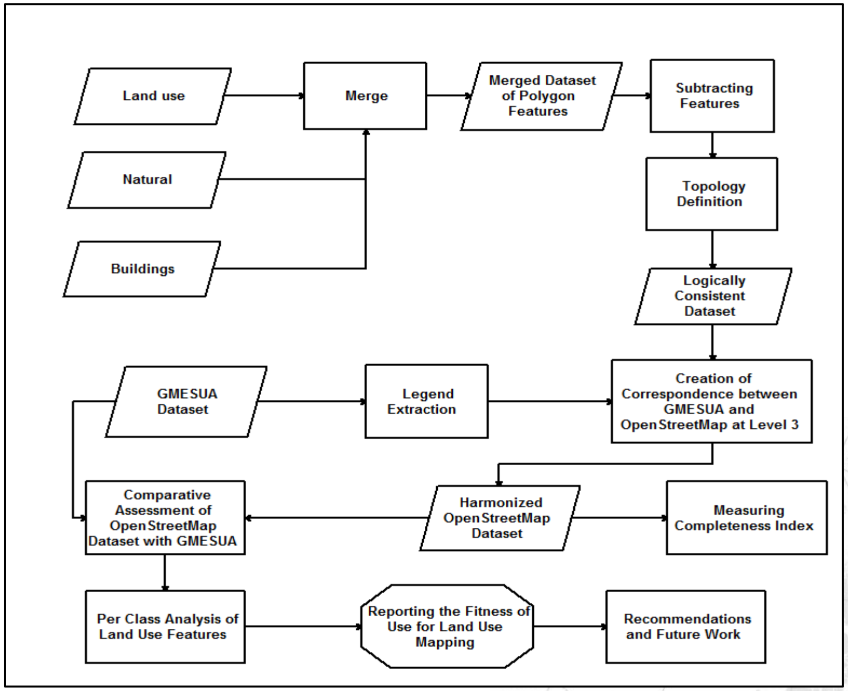
Overall accuracy = 86,2%

CONFUSION MATRIX (OVERALL)

Overall Accuracy:~ 86%

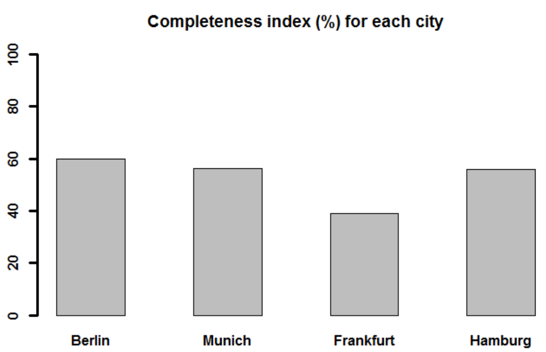
Objectives

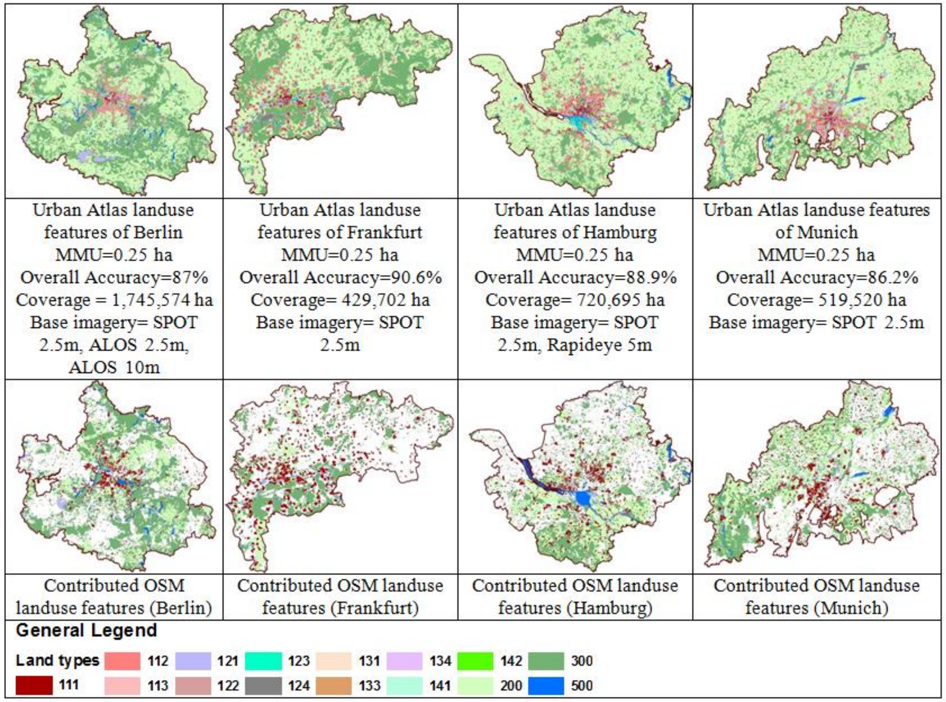
- To comparatively evaluate the quality of the contributed OSM land-use features and
- To see how reliable we could start exploiting them.



Selected areas

- Berlin,
- Frankfurt,
- Hamburg,
- Munich





Jokar Arsanjani, J., Mooney, P., Zipf, A., Schauss, A., (2015): Quality assessment of the contributed land use information from OpenStreetMap versus authoritative datasets. In: Jokar Arsanjani, J., Zipf, A., Mooney, P., Helbich, M., (eds) OpenStreetMap in GIScience: experiences, research, applications. ISBN:978-3-319-14279-1, PP. pending, Springer Press.

Thematic accuracy

Thematic accuracy

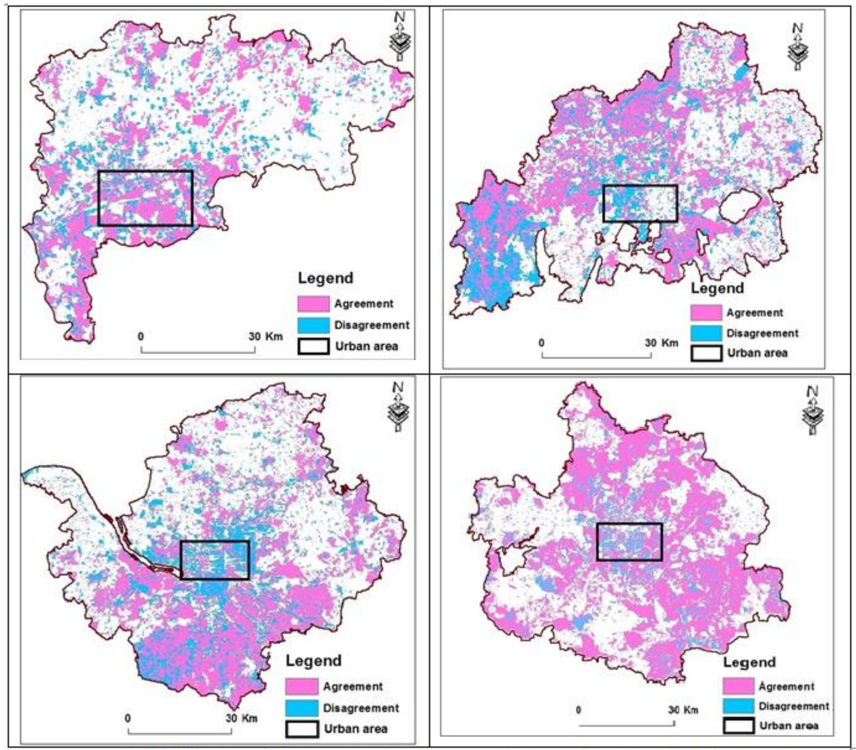
- Regardless of the degree of data completeness

City	Overall Accuracy
Frankfurt	76.5%
Hamburg	63.8%
Berlin	75.9%
Munich	67.1%

Overall Accuracy=63.86%	Contributed LU (OSM) in Hamburg																
Kappa Index=0.408	111	112	113	121	122	123	124	131	132	133	134	141	142	200	300	500	Total	User's Accuracy (%)
Reference Data (UA) 111	62817.2	0	0	2095.6	66.8	0	0	4.8	0	80.8	11.6	27512.8	88	891.2	345.2	1750.4	**95664.4**	65.66
112	396666.8	0	0	1380.4	64.8	0	0	1.6	0	92.8	81.2	33150.8	751.6	16344.8	10155.2	1612	**460302**	0.00
113	7202.8	0	0	177.6	0	0	0	0	22	6.8	0.4	1358	14.4	7604	3494.4	69.6	**19950**	0.00
121	37348.4	0	0	94455	410	0	0	414	652	746.4	849.2	24061.6	214	16968.4	5110.4	2659.2	**183888.4**	51.37
122	12313.2	0	0	3163.6	2979.6	0	0	18	2.4	24.4	53.2	8427.2	71.2	3479.6	6574.4	3153.2	**40260**	7.40
123	2.8	0	0	19478	6.4	0	0	0	1.2	6.4	116.8	474	3.6	3.2	96.4	5222.8	**25412**	0.00
124	0	0	0	751.2	0	0	0	0	0	11.6	0	2180	4.8	1.2	1782.8	0.4	**4732**	0.00
131	78.4	0	0	621.2	0	0	0	9885.2	3364.4	2.8	0	222	134.8	704.8	988.4	190.4	**16192.4**	61.05
132	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**0**	-
133	1697.6	0	0	883.6	7.2	0	0	40.4	0	604	177.2	365.2	0	563.2	263.6	406.4	**5008.4**	12.06
134	2442.8	0	0	1012.8	13.6	0	0	0	15.2	50.8	107.6	348.4	30	254.4	328.4	81.6	**4685.6**	2.30
141	7983.2	0	0	885.2	47.6	0	0	30.4	25.2	37.6	38.8	25317.2	856.4	420.8	11145.2	1876.8	**48664.4**	52.02
142	6192.8	0	0	906.4	30.4	0	0	99.2	5.2	32	136.4	11430.8	16632	3464.8	4578.4	2555.2	**46063.6**	36.11
200	26378.8	0	0	12068	84.8	0	0	3840	2238	1019.6	960.4	164743	878.8	1271068	302746	18981.6	**1805006.4**	70.42
300	7658	0	0	4687.6	45.6	0	0	150.8	585.6	52.4	28.4	50048	175.2	22349.6	846946.4	4378.4	**937106**	90.38
500	310.4	0	0	396	2.8	0	0	564	1692	28	0.8	8150.8	84	618	4496	105036	**121378.8**	86.54
Total	**569093.2**	**0**	**0**	**142962**	**3759.6**	**0**	**0**	**15048.4**	**8603.2**	**2796.4**	**2562**	**357790**	**19938.8**	**1344736**	**1199051.2**	**147974.0**	**3814314.4**	-
Producer's Accuracy (%)	11.04	-	-	66.07	79.25	-	-	65.69	0.00	21.60	4.20	7.08	83.42	94.52	70.63	70.98	-	-

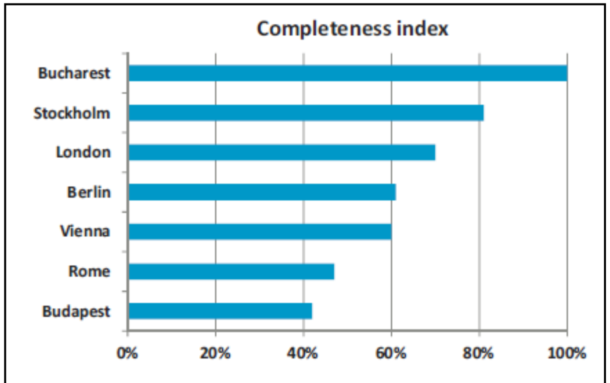
Urban/Rural

- Despite road features, relatively similar rate of contribution in rural and urban patterns,
- This could be due to the fact that dense features in cities do NOT let users to map land use features,

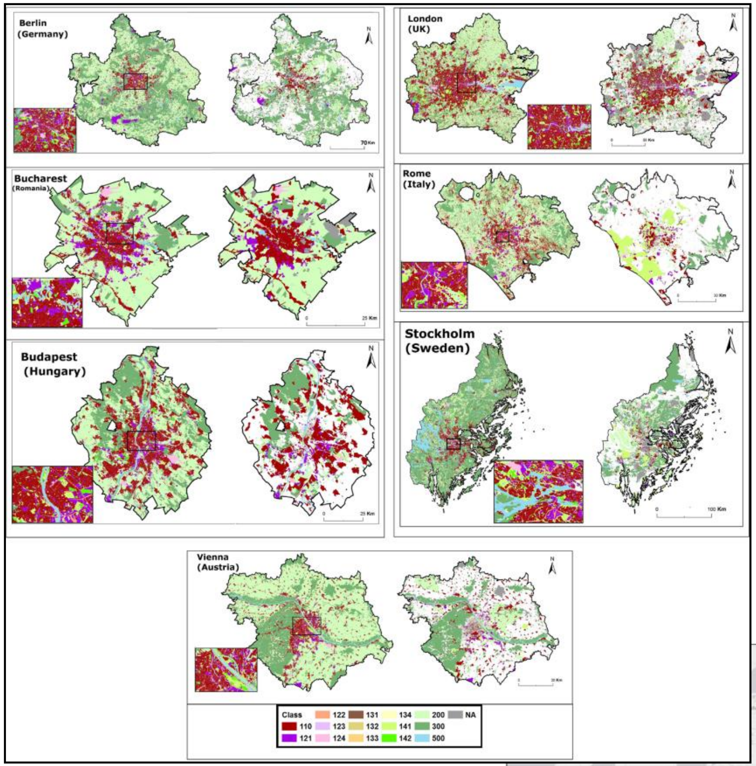


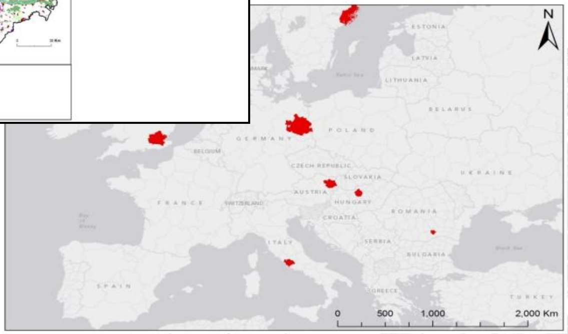
spatial distribution of agreement and disagreement between OpenStreetMap landuse features and GMESUA dataset for Frankfurt (top-left), Munich (top-right), Hamburg (down-left), and Berlin

European cities



The degree of completeness for each case study.





Thematic accuracy

- Substantial agreement for some classes (Urban fabrics, water bodies, forest) in general and depending on the city, some other classes are highlighted.

User's and producer's accuracies based on a per class analysis of contributed land use features to OSM compared to GMESUA land use datasets.

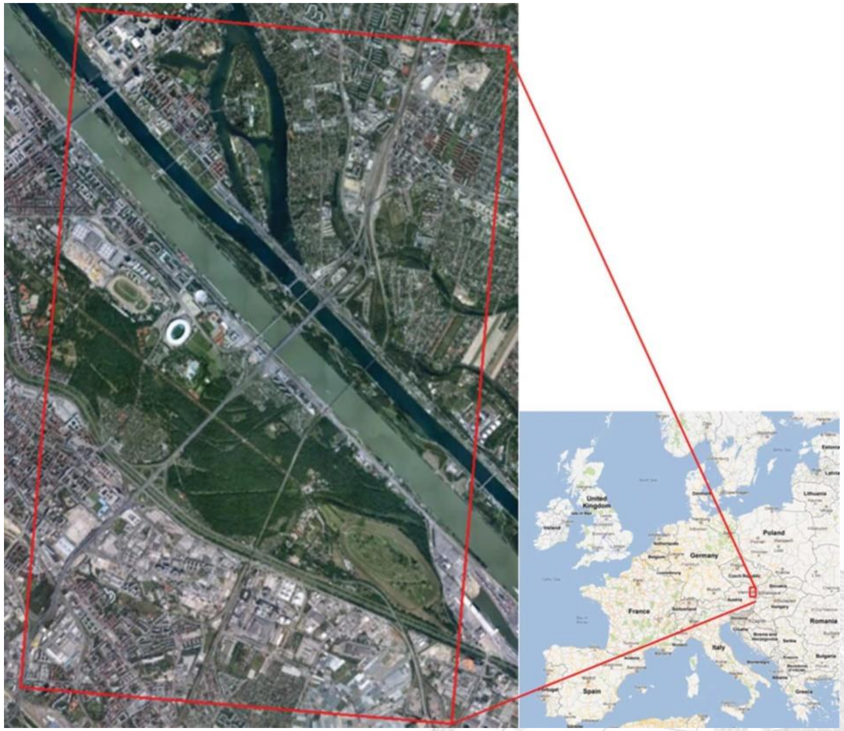
Land type	Accuracies	Metropolitan area						
		Budapest	Rome	Vienna	Berlin	London	Stockholm	Bucharest
Urban fabrics [110]	Producer's accuracy	74.12	71.31	69.06	70.09	79.00	74.59	63.83
	User's accuracy	96.73	82.94	92.14	77.59	93.53	66.68	86.84
Industrial, commercial, public, military and private units [121]	Producer's accuracy	64.65	68.07	57.05	50.59	57.31	67.42	60.79
	User's accuracy	64.48	58.32	62.36	30.93	60.43	53.16	44.36
Road and rail networks and associated land [122]	Producer's accuracy	46.26	32.45	41.45	67.73	24.95	26.79	37.84
	User's accuracy	8.43	8.42	8.08	2.22	2.80	6.70	7.90
Port areas [123]	Producer's accuracy	0.00	0.00	0.00	70.53	1.36	2.42	–
	User's accuracy	0.00	0.00	0.00	8.33	0.13	0.85	–
Airports [124]	Producer's accuracy	94.88	81.42	–	–	83.92	77.72	92.67
	User's accuracy	83.07	98.53	0.00	0.00	76.06	90.48	92.18
Mineral extraction and dump sites [131]	Producer's accuracy	70.37	71.94	60.85	76.96	50.41	77.63	–
	User's accuracy	38.30	51.66	69.17	25.93	66.72	68.62	0.00
Construction sites [133]	Producer's accuracy	6.06	5.91	5.56	6.02	7.91	25.60	0.66
	User's accuracy	4.48	1.93	3.59	2.74	9.45	31.18	0.08
Land without current use [134]	Producer's accuracy	0.00	1.77	0.73	0.14	2.08	0.00	0.00
	User's accuracy	0.00	0.95	0.85	1.18	1.39	0.00	0.00
Green urban areas [141]	Producer's accuracy	29.72	3.93	49.66	7.78	44.12	4.35	76.81
	User's accuracy	29.64	54.16	37.68	26.35	47.13	15.94	58.97
Sport and leisure facilities [142]	Producer's accuracy	63.49	78.97	69.29	67.81	68.93	79.84	78.13
	User's accuracy	19.74	50.78	50.20	38.47	58.98	61.77	32.67
Agricultural areas, semi-natural areas and wetlands [200]	Producer's accuracy	92.85	85.89	92.56	94.52	85.48	58.54	92.04
	User's accuracy	59.90	16.67	80.97	44.93	73.29	71.22	92.76
Forests [300]	Producer's accuracy	87.31	73.31	93.35	55.51	76.76	89.50	91.53
	User's accuracy	95.68	89.50	97.62	87.23	92.57	82.31	92.95
Water [500]	Producer's accuracy	90.48	83.79	87.90	27.91	84.63	91.78	91.82
	User's accuracy	97.18	85.23	89.39	55.23	72.87	86.54	78.90

Jokar Arsanjani, J. & Vaz, E. (2015): An assessment of a collaborative mapping approach for exploring land use patterns for several European metropolises. International Journal of Applied Earth Observation and Geoinformation. DOI: 10.1016/j.jag.2014.09.009

HOW TO ENRICH THE INCOMPLETE DATA

Study site and data

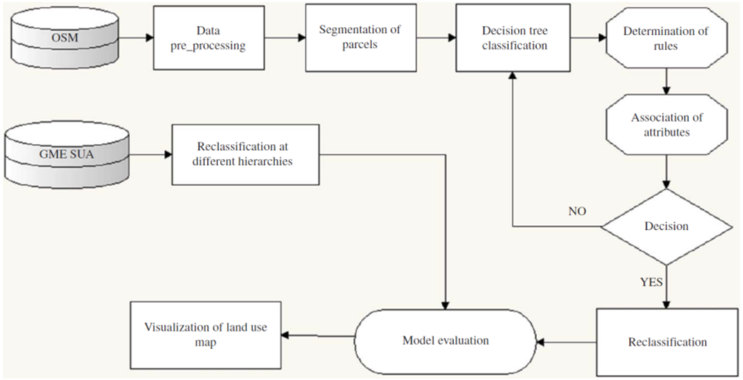
- An exemplary area within the central part of the city of Vienna, Austria, was selected.
- The reason for selecting Vienna is that it has attracted a significant amount of contributions according to a query to OSMatrix.
- The selected area of interest (AOI) covers a diverse landscape so that several LU features can be detected, including water bodies, agricultural areas, urban fabrics, and artificial surfaces. Moreover, the GMESUA data for the AOI has already been generated and released. The AOI contains 12 major land types within a 32-km^2 area.

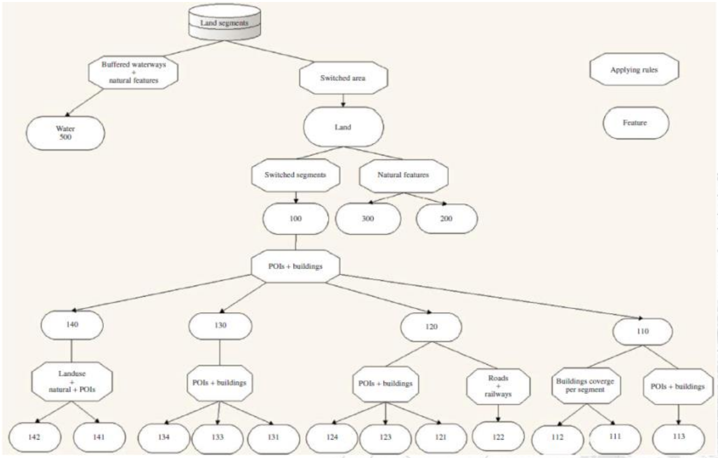


Methodology

The following tasks were followed to carry out this research:

- Data pre-processing
- Segmentation of land (spatial units)
- Determination of land attributes
- Hierarchical GIS-based decision tree approach
- Suitability analysis (texture-variability analysis, relative richness, fractal dimension analysis, Kappa_no, Kappa_location)





Results

- The output land use maps were evaluated versus GMESUA in two ways: a) a statistical analysis of texture, b) applying an error matrix of classification.
- The computed Kappa indices for each level of classification show that *artificial surfaces (100)*, agricultural+*semi-natural* areas+ *wetlands* (200), *forests* (300) and *water* (500) can be extracted from OSM at high degree of accuracy.

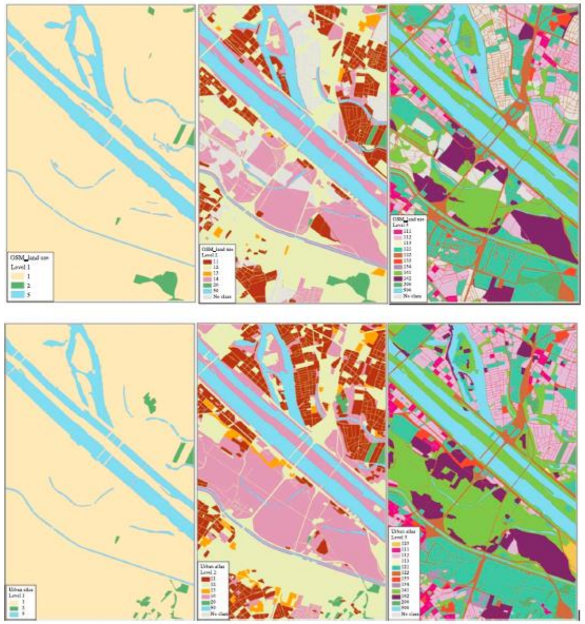
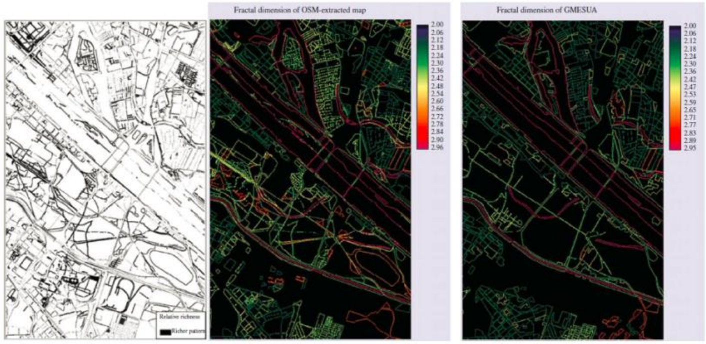


Table 2: The computed Kappa indices for each level of classification

Level	$Kappa_{Location}$	$Kappa_{No}$
1	90.64 %	81.47 %
2	78.69 %	67.38 %
3	75.58%	64.79 %




International Journal of Geographical Information Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

Toward mapping land-use patterns from volunteered geographic information

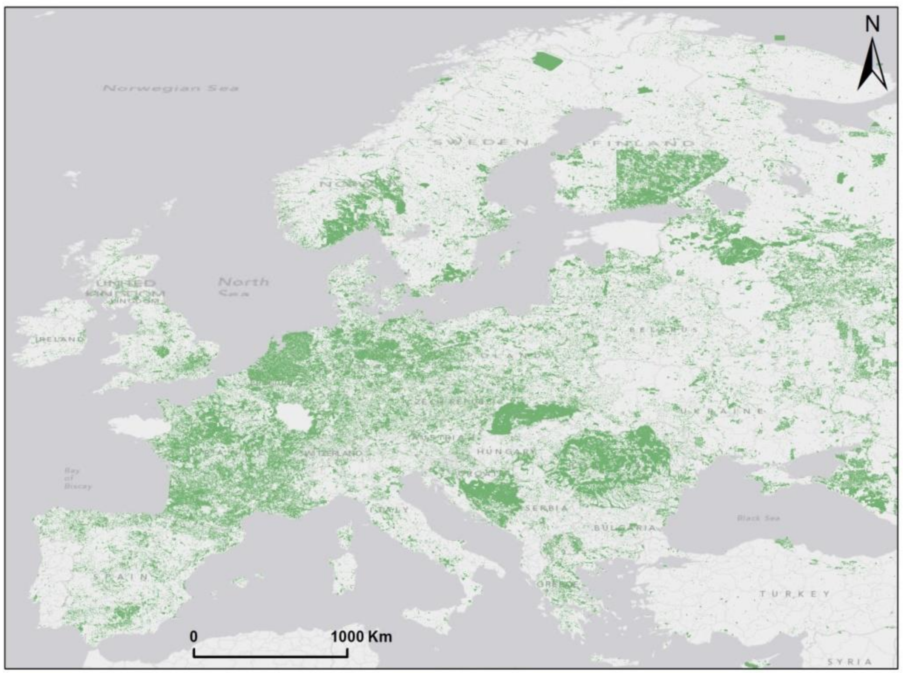
Jamal Jokar Arsanjani[a], Marco Helbich[a], Mohamed Bakillah[a], Julian Hagenauer[a] & Alexander Zipf[a]

[a] Institute of Geography, GIScience Group, Heidelberg University, Heidelberg, Germany

Published online: 19 Jun 2013.

European scale: Completeness

- Class A is those countries that > 50% mapped, while Class B is those < 50% mapped.
- 27% of whole Europe is mapped, while 36% of west
- Europe is mapped.

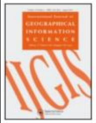


Jokar Arsanjani, J., Vaz, E. & Bakillah, M. Mooney, P. (2014): Towards initiating OpenLandMap founded on citizens' science: The current status of land use features of OpenStreetMap in Europe, in proceedings of the 17th AGILE Conference on Geographic Information Science, Castellon, Spain.

Country	Total Area (km2)	Mapped Area (km2)	Completeness (%)	Class
Bosnia & H.	51,209	49,495	96.6	A
Slovakia	49,035	43,698	89.1	A
Netherlands	37,354	30,818	82.5	A
Belgium	30,528	19,221	63.0	A
Romania	238,391	138,737	58.2	A
Luxemburg	2,586	1,426	55.2	A
France	548,500	296,833	54.1	A
Germany	357,114	190,851	53.4	A
Liechtenstein	160	65	41.2	B
Macedonia	25,713	9,432	36.7	B
Czech R.	78,867	28,728	36.4	B
Croatia	56,594	17,591	31.1	B
Andorra	468	144	30.9	B
Poland	312,685	88,489	28.3	B
Austria	83,945	22,764	27.1	B
Denmark	43,094	11,610	26.9	B
Switzerland	41,277	10,803	26.2	B
Cyprus	9,251	2,422	26.2	B
Slovenia	20,273	5,240	25.8	B
Finland	338,419	86,569	25.6	B
Montenegro	13,812	2,916	21.1	B
Spain	505,992	106,131	21.0	B
Greece	131,957	27,181	20.6	B
Great Britain	242,900	46,366	19.1	B
Lithuania	65,300	12,108	18.5	B
Kosovo	10,908	2,004	18.4	B
Norway	386,224	61,706	16.0	B
Moldova	33,846	5,410	16.0	B
Malta	316	48	15.4	B
Hungary	93,028	14,198	15.3	B
Serbia	88,361	11,481	13.0	B
Bulgaria	110,879	14,362	12.9	B
Sweden	441,370	56,657	12.8	B
Italy	301,336	38,024	12.6	B
Ukraine	603,500	68,735	11.4	B
Belarus	207,600	22,968	11.1	B
Ireland	70,273	4,965	7.1	B
Portugal	92,090	3,919	4.3	B
Albania	28,748	897	3.1	B
Iceland	103,000	1,687	1.6	B

Lessons learned

- There is a huge potential for regional applications, given the high degree of accuracy that these products offer.
- The considerable completeness variation confirms the heterogeneity nature of VGI across space.
- Potential of applying data mining techniques for mapping land use features from the other given features.
- logical consistency perspective
- From a thematic quality perspective, the contributed features have, except for Rome (fair), a moderate rank of Kappa indices and "substantial" to "very substantial" overall accuracies are achieved.
- Per-class analysis demonstrates that classes such as U*rban fabrics [110], Airports [124], Mineral Extraction and Dump Sites [131], Sports and Leisure Facilities [142],Agricultural + semi-natural areas + wetlands [200], Forests [300], and Water Bodies [500]* have the highest accuracies and, therefore, these features could be highly exploited.
- Reference dataset:::Is GMESUA the best reference?
- Temporality issues and changes over time



International Journal of Geographical Information Science
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tgis20

Mining urban land-use patterns from volunteered geographic information by means of genetic algorithms and artificial neural networks
Julian Hagenauer [a] & Marco Helbich [a]
[a] Institute of Geography, University of Heidelberg , Berliner Straße 48, 69120 , Heidelberg , Germany
Published online: 14 Nov 2011.

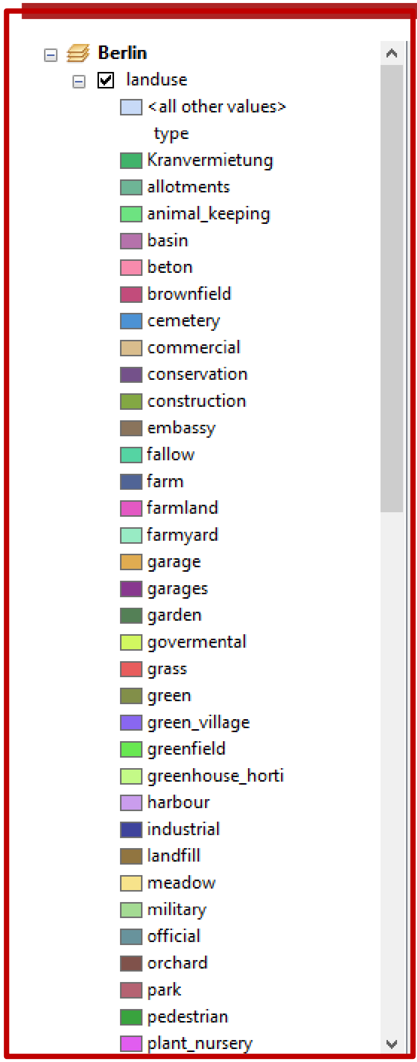
Vaz, E. & Jokar Arsanjani, J. (2014): Crowd sourced mapping of land use in urban dense environments: An assessment of Toronto. The Canadian Geographer Journal (accepted).

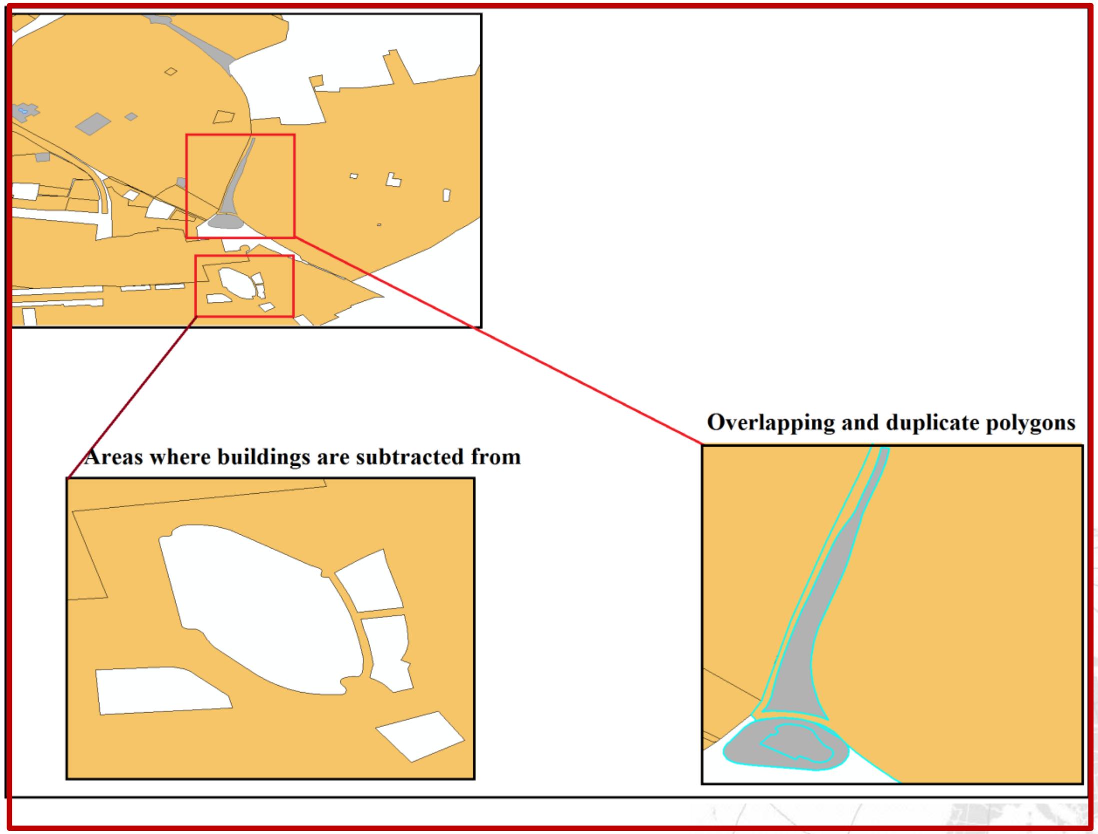
FUTURE (ONGOING) RESEARCH

Internal quality evaluation (2)

- **Logical consistency: a very challenging issue**
- Logical consistency addresses how well topological and logical relationships between the dataset segments are defined.
- The main challenge in using landuse features is the fact that the contributions have dissimilar geometrical accuracy and frequently overlap each other (i.e., some areas are given several land types).
- Therefore, using this layer for any external application generally demands for defining topology on the features in order to clean them from overlaps and dangle errors and also build their topology.
- In landuse dataset, in some parts several polygons overlap each other, which does not let us to have a single contribution for that area and therefore several attributes are existent.
- These issues are resolved by applying topological relations between the objects.

Difficulties +

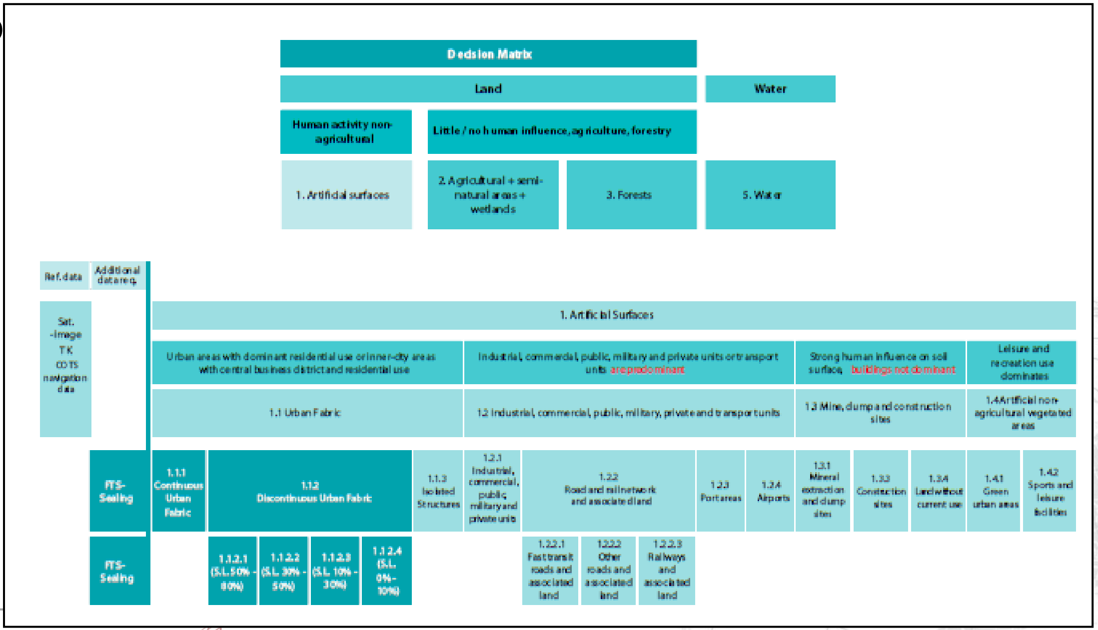




Difficulties ++

Harmonization of the datasets:

- The OSM land use features are NOT standardized. (The features do not follow the global/regional land-use classification schemes)→→ the OSM land-use features must be harmonized with the GMESUA→→ by translating the OSM land-use features into the CORINE land-cover classification scheme →→This helps to make a common landuse language and also creates a dictionary for translating the contributions of individuals to



Conclusions

- From a positional and attribute perspective, the contributed features have in general moderate rank of Kappa indices and their overall accuracies are between 63% and 77%, whilst the GMESUA datasets also have overall accuracies below 90% and therefore the computed accuracies are considerable.
- Per-class analysis of the landuse types shows that in general the *continuous urban fabrics* [111] and *agricultural+semi-natural areas+wetlands* [200], *forests* [300], and *water bodies* [500] have the highest accuracies (>80%) and therefore, their accuracies prove that these features could **highly be exploited**.
- **It is not guaranteed** that GMESUA datasets account for the best reference datasets to evaluate the OSM landuse features based on them whereas there are some concerns on the accuracy of GMESUA datasets. This might have caused some errors and misclassifications such as follows: a) the accuracy of the GMESUA datasets varies between 83% and 90%, b) according to their metadata, archived images of 2005 until 2010 have been used for landuse mapping and this could have caused a major source of disagreement whereas the OSM landuse contributions have been mainly lately (majority after 2009) given.
- The OSM landuse features message a promising alternative source of landuse mapping independent from applying computational signal processing techniques coupled with expert knowledge of land. Certainly the longer the lifetime of OSM becomes, the more contributions are collected and high accuracy landuse maps could be retrieved.

Enriching strategies

- Data mining and knowledge discovery
- Mapping events
- Importing available local datasets (Slovenian case)
- General awareness
- Parallel workshop VALID-LAND→ enriching OSM by photos initiatives
- What other issues?

Thank you for your attention!

Questions, comments!!