Slide 1

Good afternoon,

Today I am going to present you the work that we are doing at the Spatial Epidemiology Lab and the Royal Military Academy in Brussels,

to improve the accuracy of built-up maps in developing regions – more specifically Sub-Saharan Africa, by making use of the increasing amount of available geographic data sources.

Slide 2

As you may know, the first wave of open acess satellite imagery, with missions such as MODIS or MERIS, enabled the production of a serie of global built-up maps at a coarse resolution such as: - the Global Urban-Rural Mapping Project in 2006 - or Globcover in 2010 ;

Here you can see an example of this datasets on the city of Windhoek, in Nambia.

Slide 3

Later, the Landsat catalogue has been made freely available. And researchers took advantage of the opportunity to create a new serie of global built-up maps at a finer spatial resolution, for example the GHSL and the HBASE datasets.

As you may know, studying the urbanization process and its impacts on developing regions is one of the main application of these datasets, because it is where the stakes are actually located.

Slide 4

However, that is also where the accuracy and the reliability of such maps are the lowest.

This is due to many issues:

- First, of course, there is a lack of reference data to support the training or the validation of the classification models ;
- But also imagery-related issues such as the spectral confusion that occurs between the spectral signature of the built-up and the bare soil.

This is especially problematic in arid regions and where buildings are built with nearby natural materials, such as here in Gao, in Mali.

Slide 5

Additionnally, classification algorithms suffer from the inherent spectral heterogeneity of the built-up environment.

In the past, researchers tried to identify a spectral definition of a built-up pixel. However the only finding was that there was not any consistency depending on the location.

Slide 6

First, inter-urban heterogeneity is observed, because of socio-economic, historical, cultural, or environmental differences among the cities of the world.

Here you can see an example of three urban mosaics, on the same scale, in three sub-saharan african cities.

This means that a classification model that works on a given city cannot be applied on another one as it is. This also means that a classification method can work, in Europe for example, but not in Africa.

Slide 7

There is also intra-urban heterogeneity inside a given city, caused by variations of the urban mosaic in terms of building materials, building density, vegetation density, etc.

For example here in the wity of Windhoek, in Namibia.

Slide 8

We choose to investigate potential improvements to large-scaled built-up mapping by taking advantage of two major trends:

– the increasing availability of different types of open-access satellite imagery, more specifically Optical and SAR imagery ;

– the growing amount of geographic information in crowd-sourced projects such as OpenStreetMap.

Slide 9

As you may know, one of the biggest issue in large-scaled classification of remote sensing imagery – at least if you want to use supervised algorithms – is the collection of training samples.

They can be manually digitized to ensure the spatial and spectral representativeness of the sample, but it is hugely time-consuming.

They can also be derived from existing land cover databases – however in this case we are havily dependent on the datasets quality.

So we choose to make use of the growing amount of data available in OpenStreetMap to collect both built-up and non-built-up samples. In the first part of the presentation, we will investigate if such sampling strategy can be reliable in developing regions such as Sub-Saharan Africa.

In the second part of the talk, we will investigate the combined use of both optical and SAR data to tackle issues such as the spectral confusion between built-up and bare soil.

Slide 12

The first issue that can be encountered when using OSM is the spatial heterogeneity of data quality and availability ; which is caused by two phenomenons :

– firstly, the pride of place : users are more likely to contribute where they live or when they have an economical interest (for example in the case of administrative or commercial contributions).

– and the Digital Divide, which is caused by an inequel access to education and Internet.

This lead to, in the case of OSM:

– 10 times more information (in bytes) in Europe than in Africa ; – or also 2 times more information in Germany than in the whole Sub-Saharan Africa region.

This is as of March 2018...

Slide 13

I specify March 2018, because it is rapidly changing.

This graphics shows the per-continent evolution of the bytes of information in the OSM database (on a logarithmic scale).

It reveals that the amount of information available in continents such as Africa or Asia are catching up with Europe and North America. As a matter of fact, Africa is the continent where contributions are increasing at the highest rate since 2014.

This is what makes this study possible today.

Slide 14

Our study will be focused on 10 different case studies in Sub-Saharan Africa, with various environmental and population patterns.

Slide 15

Some of you may be already familiar with data model of OSM, but as a remainder:

OSM follows a very simple data model, with each object described at least on key/value pair.

Our objective is to collect both built-up and non-built-up training samples. So our first idea was to use building footprints as built-up samples, and specific land use, natural or leisure features as non-built-up samples.

Slide 16

As you can see, the amount of building footprints available in some cities is enormous and can be more than sufficient to train a supervised classifier.

Slide 17

Likewise, large urban areas benefits from a sufficient amount of information regarding non-built-up areas such as vegetated areas, coastlines, parks, wet lands, etc.

Slide 18

However, that is not the case in all urban areas. In fact, most of the time there is a lack of information either for built-up areas or non-builtup areas.

For example, smaller urban areas such as Gao, Katsina or Windhoek does not benefit from a sufficient amount of building footprints. Additionnally – most of the time, mapping campaigns focuses on specific neighborhoods or informal habitat. That can lead to a training dataset which is not representative of the urban heterogeneity in a city.

Slide 19

To collect more training samples, we performed two spatial analyses on the road network – which is the most exhaustive feature in OSM.

For built-up samples, we used the concept of urban blocks – defined as the polygons shaped by the intersection of the road segments. By considering only a specific type of roads (such as residential roads) and by avoiding the largest polygons which are the most likely to contain empty spaces or mixed pixels, we can derive actual information on built-up areas.

Slide 20

Regarding non-built-up samples, this is more problematic because OSM is focused on the urban environment. However, we can use the built-up information to identify the areas which are the less likely to be built.

We used the concept of urban distance – which is, in this case, the distance from any road or building, to randomly collect training points in the most remote areas which are likely to be non-built.

Slide 23

To assess the reliability of the strategy, we performed three different classifications on a Landsat 8 scene for each case study.

– the first one using reference hand-digitized polygons as training samples ; – the second one using first-order OSM features such as building footprints or non-built-up objects such as parks, vegetated areas, etc. – the last one using the previous features plus the urban blocks and the urban distance.

Slide 24

This table shows the validation metrics for the three classification schemes.

As you can see, using only first-order OSM features such as building footprints and non-built-up objects leads to very low accuracies when the amount of available samples is not sufficient. This is the case for Antananarivo, Chimoio, Katsina or Windhoek. In Johannesburg, most of the building footprints were located in the city center. This led to high rates of misclassifications in the other parts of the city.

However, adding urban blocks and urban distance allows for better accuracies in most of the cases.

And finally, the results shows that the classification performance can reach similar scores than when using hand-digitzed training samples.

Slide 27

Of course, the performance of the classification is heavily dependent on the number of training samples, as you can see in this graph.

Slide 28

Regarding the combined use of optical and SAR data:

as you may know, they are complementary in theory:

Optical data separates well vegetation and built-up, and SAR data separates well bare soil and built-up.

This should increase the classification performance, especially in arid regions and when buildings are made of nearby natural materials.

Slide 29

This lead us to this classification scheme, which is simpler than it looks:

We have two types of input imagery: SAR and optical, from which we extract GLCM textures and thematic indexes such as the NDVI.

We extract training samples from OSM to train a classification model based on the RF algorithm.

As such, the fusion between Optical and SAR occurs at the pixel-level during the classification step.

Slide 30

This table presents the F1-score of three different classifications:

– one when using only optical data ; – one when using only sar data ; – and one when using both.

As you can see, combining optical with SAR data does increase the classification performance in most of the case studies. This is especially true in urban areas located in arid regions such as Windhoek or Gao.

Slide 32

However, the benefit is more visible when assessing the classification performance in specific land covers.

Here, in Windhoek, including SAR data increase the F1-score by only 3 points. But as you can see – land covers are affected differently. As a matter of fact, it increases the accuracy in bare soil areas by 16 points.

Slide 33

To sum up…

We found that the complementarity between optical and sar data does increase the classification performance in sub-saharan africa – even with very simple pixel-level fusion methods. This is especially true in arid regions.

We also found out that OSM data can be used to automate the training data collection step – provided that (1) spatial analyses are conducted on the raw data, and that (2) a sufficient amount of data is available in a first place.

The fact that the performance of our method is heavily dependant on the amount of available data in OSM makes these results promising, because more and more geographic information are added in OSM each day. In fact, 70% of the OSM objects used in this study were not available three years ago.

Slide 34

This is the end of this presentation, many thanks for your attention !

If you want to see the code, it is available on github (mainly for the OSM part).

If you want to explore our results, a web map is available on the website of our research project.

(If you want to find flaws in it !)

Thanks ! Any questions ?