

## Chapter 3

# Towards automated regional data compilation in the IDE-GSM

Souknilanh Keola\*

### Abstract

This chapter traces the history of regional data compilation in the Geographical Simulation Model (IDE-GSM), a computational general equilibrium model based on spatial economics. Regional data compilation started with manual inputs from ground-based surveys and censuses, before evolving to partly employ remote-sensing data and machine-learning techniques. The application of OpenStreetMap (OSM) data has been explored, but it is yet to be used to generate regional data. None of these techniques is absolutely superior now or in the foreseeable future. Various combinations of the aforementioned approaches on a case-by-case basis are likely to continue to be workable solutions.

**Keywords:** Regional data compilation, Remote sensing, OpenStreetMap (OSM).

### Introduction

The number of regions in the Geographical Simulation Model (IDE-GSM) started from several hundred in Version 1 (Kumagai et al. 2008), increased continuously as the geographical coverage expanded, and had reached more than 2,000 by 2013 (Kumagai et al. 2013). The IDE-GSM requires two types of data as inputs to run the simulation: (i) regional data (city data) and (ii) route data connecting all regions in several modes (land,

---

\* Deputy Director, Economic Geography Studies Group, Development Studies Center, Institute of Developing Economies, JETRO (IDE-JETRO).

sea, air, and railway). This chapter focuses on the regional data, traces how this task was initially conducted and how it evolved over time, and discusses possible ways forward.

## 2. Regional data in the IDE-GSM

There are two types of regions in the IDE-GSM. The first is regions or cities where people live and work for firms located there. Theoretically, all people and firms in each region live and are located in the center of each region. These regions are labeled habitable within the database. The second type are regions that exist in the simulation system solely for the purpose of making the road or other modes of logistic connectivity more realistic. These regions are labeled as inhabitable in the model. The only information required for inhabitable regions is their location. The regional data discussed in this chapter refer only to those compiled for habitable regions. Unless otherwise stated, region refers to a habitable region.

The regional data of regions in the IDE-GSM are largely divided into population, employment, and value-added data. The population is simply the total population of each region. Employment and value-added need to be compiled by industries such as agriculture, services, and five or more manufacturing industries. The next section traces how the compilation of these regional data in the IDE-GSM has evolved since 2007.

## 3. Compilation of regional data in the IDE-GSM

Regional data compilation in the IDE-GSM started with manual computation and input based on surveys or censuses. Remote-sensing data, especially nighttime light (NTL) and land cover data, have been used to generate regional data since around 2012, for countries where official data such as economic censuses are not available at all (Kumagai et al. 2012). Although still in the experimental stage, machine learning is used to create regional data for a selected country in 2020. In this fiscal year, OpenStreetMap (OSM), an open online spatial database created by volunteers around the world, is examined as a new way to compile regional data. The remainder of this section details the approaches.

### 3.1. Surveys and censuses

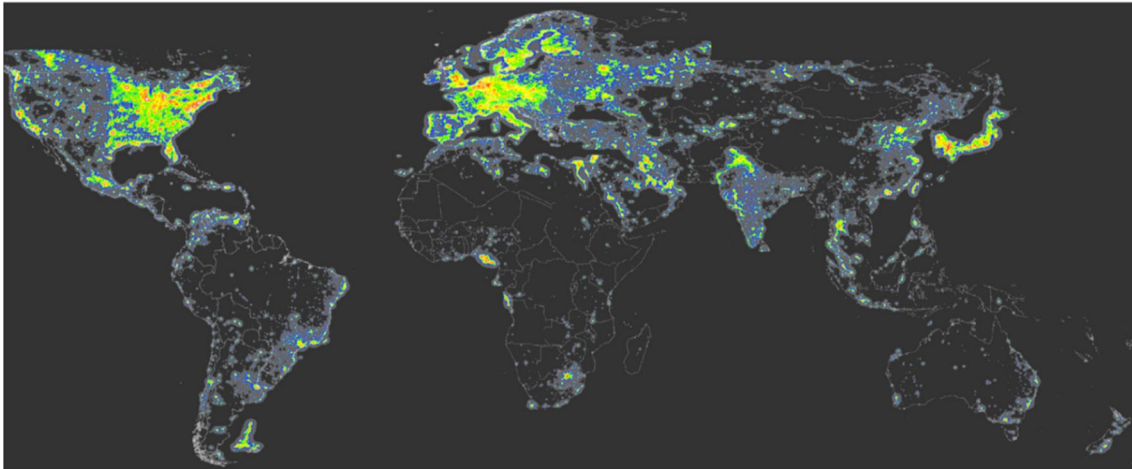
The library of the Institute of Developing Economies (IDE) houses a vast collection of statistical publications from developing countries. Compilation of regional data from these publications was, therefore, a natural choice at the beginning. For example, once the base year is fixed, the regional population is compiled from the population census closest to the base year. Value-added data are compiled in a similar way based on various economic surveys and census data. Administrative, survey, and census data are generally considered ground truth in social economic studies. If the required data are available, the regional data generated from them are considered to be of the highest quality.

The biggest drawback of this approach is manifold. The first is the availability of data. Population censuses are often conducted only once in a decade, even in developed countries. The interval becomes much longer in vulnerable states under many types of internal conflicts. Unfortunately, this is often the case in many developing countries. Even when data exist, they may differ from the regional classification used in the IDE-GSM base year. The second is consistency across countries. Industrial classification differs not only among countries but also within the same country over time. Furthermore, the quality varies greatly across countries. Third, and probably the most critical drawback of this approach is its financial and time cost. Population and economic censuses are generally very expensive. The number of region mid-scale countries is usually close to 100, while it is several hundred for large countries such as China, India, and Indonesia. Compiling value-added by, for example, seven industries for every region in a mid-scale country requires a substantial amount of time. As the coverage of the IDE-GSM expanded to cover more countries in Asia, Europe, America, and recently Africa, the compilation of regional data based on surveys and censuses quickly became difficult to sustain.

### 3.2. Remote-sensing data

Many studies in the field of economics have started using remotely sensed artificial nighttime light data to quantify the scale regional economy (Chen and Nordhaus 2011; Doll et al. 2006; Ebener et al. 2005; Elvidge et al. 1997; Henderson et al. 2012; Sutton and Costanza 2002; Keola et al. 2015; Tanaka and Keola 2017). The IDE-GSM first generates regional data for Myanmar, whose regional data were mostly unavailable at that time, using the Global Defense Meteorological Satellite Program - Operational Line-Scan System (DMSP-OLS) Nighttime Lights Time Series 1992-2013 (Kumagai et al. 2012). The advantage of the NTL is its very high spatial resolution (Figure 3-1), which, based on

**Figure 3-1:** A global view of nighttime lights in 2012



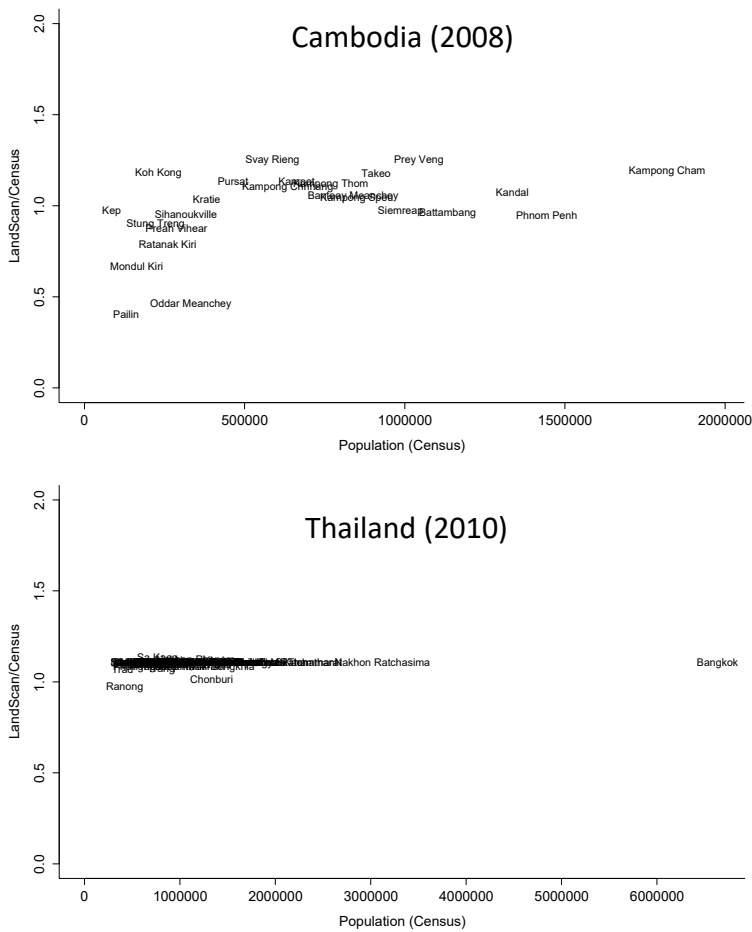
**Source:** DMSP-OLS.

DMSP-OLS, is 30 arc sec or 0.86 square kilometers at the Equator. The freely downloadable data are provided as an annual average; therefore, in principle, the annual scale of economic activities of virtually any region in the world can be generated from these data between 1992–2013. Version 5 is expected to be published in mid-2021, which is expected to extend the temporal coverage in recent years. Since the update is expected to continue in the foreseeable future, NTL based on DMSP-OLS is likely to remain one of the major sources of remote-sensing data to study regional economies, especially in regions with insufficient ground-based data for many years.

Application of remote-sensing data in the IDE-GSM went well beyond NTL. The second remote-sensing data used to generate regional data in the IDE-GSM is LandScan, a global population grid data available since 2001.

Figure 3-2 provides a comparison of the population aggregated from LandScan and ground-based censuses in Cambodia and Thailand. Population censuses are often carried out by questioning the heads of villages or settlements. In many cases, they represent reported registered inhabitants. On the other hand, LandScan estimates the average day-time population with several high-resolution remote-sensing data, including sub-meter satellite imagery. It is not surprising that both figures do not perfectly agree. In Cambodia, estimation by LandScan fits rather well with census data in major regions, such as Phnom Penh, Siemreap, Battambang, etc. The ratio between the population estimated by LandScan and Census was approximately 1 (Figure 3-2, top). However, in remote provinces such as Pailin or Oddard Meanchey, LandScan estimates the population to be much

**Figure 3-2: LandScan VS census in selected countries**



**Source:** Computed from LandScan and Global Administrative Unit Layers (GAUL). Censuses are from Cambodia Statistical Institute and National Statistical Office of Thailand.

less than the official statistics. It should be noted that it is not obvious which is correct, and both can be wrong. However, LandScan estimation is based on a consistent algorithm across countries. On the other hand, estimation by LandScan fits better with official statistics in Thailand (Figure 3-2, bottom). However, this ratio was still not 1. Except for the two provinces, LandScan predicts a larger population than official statistics in Thailand. Temporal migration of factory workers from poorer to richer regions is common in this region within or across countries. One possible explanation would therefore be that since poorer regions tend to send more people to work in richer regions, they often have more registered than actual inhabitants in Cambodia. On the other hand, all LandScan data predict all regions in Thailand to have more population than official data, because of mass migration workers coming from neighboring countries.

The last remote-sensing data used to generate regional data thus far are land cover data. There are few, if not none, economic activities that do not affect the earth's surface. In other words, the level of economic activity can be indirectly observed through changes in the earth's surface. Ground-based land-cover data are extremely limited. Some examples of such ground-based data with global coverage are the forest cover rate and crop land. Needless to say, these ground-based data are usually available only at the national level. In contrast, the gridded land cover datasets generated from the Moderate Resolution Imaging Spectroradiometer (MODIS) have a spatial resolution of approximately 500 m × 500 m. This makes it possible to generate regional land cover data for virtually any region globally.

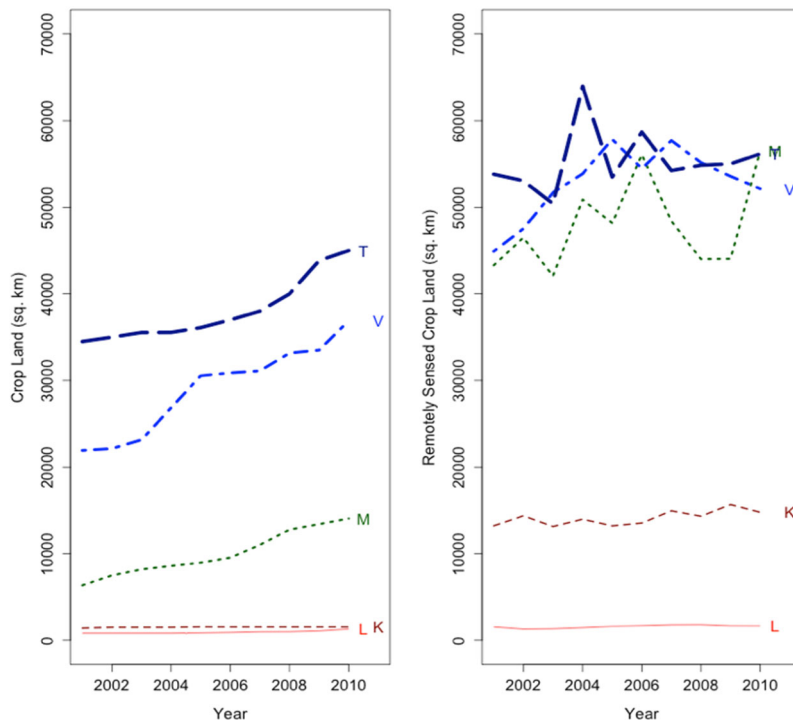
Among the many land cover data produced from MODIS, MCD12Q1 is used to divide national agricultural value-added into regional value-added for countries for which such data cannot be obtained. MCD12Q1 categorizes land cover into 16 types, from water to several types of forest, cropland, and urban land as follows:

- 0 Water
- 1 Evergreen Needleleaf Forest
- 2 Evergreen Broadleaf Forest
- 3 Deciduous Needleleaf Forest
- 4 Deciduous Broadleaf Forest
- 5 Mixed Forest
- 6 Closed Shrublands
- 7 Open Shrublands
- 8 Woody Savannas
- 9 Savannas
- 10 Grasslands
- 11 Permanent Wetlands
- 12 Croplands
- 13 Urban and Built-up
- 14 Cropland/Natural Vegetation Mosaic
- 15 Snow and Ice
- 16 Barren or Sparsely Vegetated

Figure 3-3 illustrates how remotely sensed cropland is comparable with ground-based data. On the left-hand side is cropland area extracted from the online World Bank database. Ground-based and remotely sensed cropland area in Laos largely agree. Remotely sensed cropland area in Cambodia and Myanmar is much larger than ground-based statistics. Remotely sensed cropland area in Vietnam and Thailand is relatively closer to ground-based data, but the former fluctuates substantially from year to year. Though it may be necessary to fine tune the land cover types that affect agriculture in different countries, land cover data can be a handy source to generate regional value-added of agriculture.

There are several limitations to the application of remote-sensing data to generate regional data for the IDE-GSM. For example, the 16 classifications of land cover in MCD12Q1 are quite different from those in the IDE-GSM. Cropland can be used directly to capture agriculture. Urban and built-up areas may also be used indirectly to capture the

**Figure 3-3:** Crop land change with ground-based and remote sensing data



**Source:** Ground-based cropland area is based on World Development Indicators (World Bank). Remotely sensed cropland area is computed from MODIS Land cover data set (MCD12Q1) and GAUL.

**Note:** K, L, M, V, T denote Cambodia, Laos, Myanmar, Vietnam, and Thailand, respectively.

size of services, but there is essentially no information to capture manufacturing industries. In brief, although the spatial resolution of remote-sensing data is high, the in-depth information obtainable is usually inferior to those collected on the ground. The scale of the regional economy can be captured with significant precision, but it is still not possible to disaggregate it to industries required by the IDE-GSM.

### 3.3. Machine learning

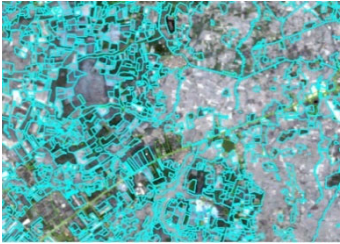
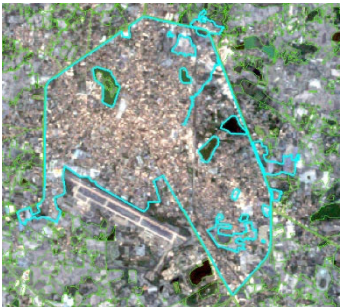

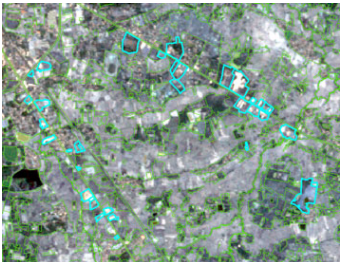
As an extension of the application of remote-sensing data in regional data compilation, in 2020, the IDE-GSM tried to apply a machine-learning technique to directly generate the type of economic activities on the ground, using (i) ground-based regional land use data and (ii) satellite imagery. This exercise was in collaboration with the Asian Institute of Technology in Thailand. The types of land use in the datasets in this exercise are as follows:

- A (Agriculture land)
- F (Forest land)
- M (Miscellaneous)
- U1 (City, Town, Commercial)
- U2 (Village)
- U3 (Institute land)
- U4 (Transportation and Communication)
- U5 (Industrial land)
- U6 (Other Built-up land)
- W (Water)

The remote-sensing data used for machine learning were satellite imagery acquired by Landsat. The Landsat program is one of the longest programs for the acquisition of satellite imagery of Earth that began on July 23, 1992. It is a joint program between the National Aeronautics and Space Administration (NASA) and the United States Geological Survey (USGS). Landsat 8, the latest of the series, was established on February 11, 2013. Table 3-1 illustrates the selected land use type (blue lines) over Landsat 8 images for Udon Thani, a province in northeastern Thailand.



**Table 3-1:** Examples of Landsat appearance by the land use class for Udon Thani, Thailand (2011)

| Land use                       | Satellite Image from ArcGIS   |
|--------------------------------|---|
| A<br>(Agricultural Land)       |   |
| U<br>(Urban and Built-up Land) | U1<br>(City, Commercial and Service)<br><br>      |
|                                | U4<br>(Transportation and Communication)<br><br> |
|                                | U5<br>(Industrial Land)<br><br>                 |

**Source:** Miyazaki (2020)

Here, the land use dataset is called a training dataset or the ground truth. The training data contain very detailed data but only for some sample sites on earth. On the other hand, Landsat 8 is remote-sensing data with global coverage. The purpose of machine learning is to train the machine to classify the type of land use from remote-sensing data. We applied Convolutional Neural Network (CNN) with U-Net architecture to the Landsat data listed above. U-Net architecture is developed at the Computer Science Department of the University of Freiburg for medical image segmentation. The algorithm was iterated

10 times and resulted in an accuracy of 0.7897. The accuracy is promising, and it is worth continuing to explore further in this direction.

The following are some of the obstacles that need to be addressed in order to compile IDE-GSM regional data to become a reality. First, more detailed land-use data are necessary. The training data used in our machine-learning exercise have nine categories, but they do not necessarily fit with what are needed for regional data in the IDE-GSM. This was the case because we had to rely on existing datasets prepared for other purposes. The training dataset with land use classes similar to the industrial classification in the IDE-GSM is ultimately necessary to build a machine-learning system that would generate regional IDE-GSM as the output.

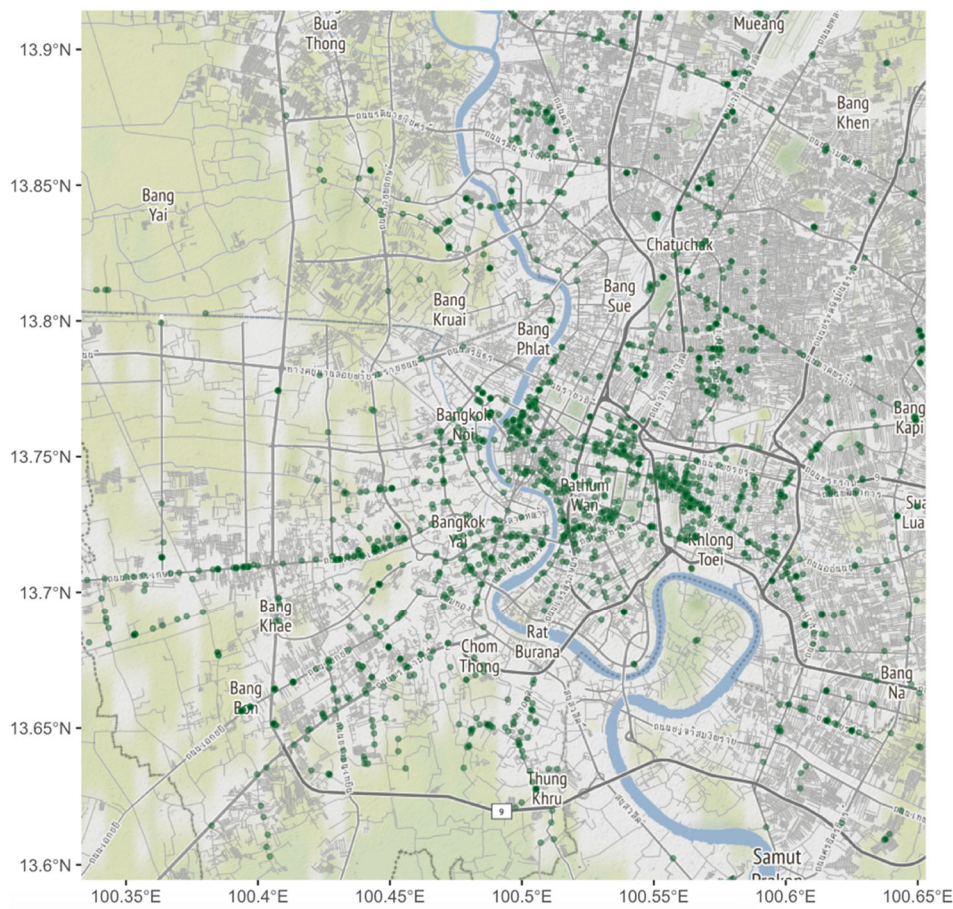
Second, the training dataset needs to cover more countries, ideally in different continents where natural environmental conditions are different. In the above example, the training dataset covers some provinces in Thailand. This kind of training data may be useful for building a machine-learning system for regions in continental Southeast Asia. However, the association between satellite imagery and training datasets may not be valid in other parts of the world, such as Europe and North America, where natural conditions and the level of economic development are different from those in Southeast Asia. Therefore, the question is how to prepare training data that cover more parts of the world. The third is how to obtain the remote-sensing data (satellite imagery) that are needed to combine with the training dataset, and generate regional data. The Landsat data were made open data in 2008, so this is currently not a problem. However, it needs to be remembered that Landsat is a domestic program funded locally by the United States, and the possibility that it would be rendered a propriety again by some future administrations is not entirely unthinkable. Other programs such as Copernicus, a European Union Earth observation program coordinated and managed by the European Commission in partnership with the European Space Agency (ESA), started to provide satellite images with a resolution comparable to that of Landsat 8. Given the coverage and continuity that the IDE-GSM aims at, the focus on open data is critical.

### 3.4. OpenStreetMap

Finally, we discuss the potential use of OSM data to generate regional data in the IDE-GSM. The OSM is a free and editable world map powered by high-resolution satellite images. It was built from scratch and is maintained by volunteers. It was released with an open-content license. Since 2004, volunteers worldwide have added spatial information, such as the shapes of roads, buildings, and points of interest (POI). As of March 2020, there were approximately 6 million registered users worldwide, of whom about 5,000 have actively contributed to updating the OSM by uploading or editing spatial data daily. We used these data for 2019. Figure 3-4 illustrates the location of shops tagged as convenience (or convenience store) in Bangkok Thailand. There are two types of data in the OSM: one representing shapes, for example, roads, buildings, and land use, and the other representing location. The former is called a polygon. As shown in Figure 3-4, the roads not only have length, but also width. This is also true for rivers, buildings, and other types of facilities. The latter is called the POI, which contains coordinates and categories. In Figure 3-4, the location of convenience stores is a category of POI plotted over many types of shapes.

The OSM has both strengths and limitations. One strength is its openness. Data in commercial search engines, i.e., Microsoft's Bing or Google Maps, are generally better, that is, more accurate, more up to date, and with wider coverage. The ample financial resources and large market powers of these global companies account for their better performance. Big and small enterprises around the world voluntarily supply their information to these search engines to increase their exposure to more potential customers. The problem is that they are not free. Furthermore, since the data cannot be downloaded in bulk, it would take a very long time, the cost would likely be prohibitively high, and sophisticated programs would be required to compile comprehensive subnational data in the IDE-GSM.

**Figure 3-4:** Location of shop tagged as convenience in Bangkok, Thailand



**Source:** Author based on osmdata package in R programming language.

On the other hand, all raw data of the OSM are available for download in bulk within approximately 48 hours. Extracting data using a required spatial unit of analysis requires much less time, money, and computational resources, but the limitation is apparently uneven coverage. The OSM data must be input by volunteers. Developed countries have more people with better internet access, higher technical skills, and willingness to contribute. Tourists, scholars, and students traveling from developed to developing countries have contributed substantially. There are several global-scale activities (e.g., the missing maps movement) that aim to increase the coverage of OSM. However, OSM coverage is still much better in higher income countries than in developing countries. This is less a problem for the IDE-GSM if the purpose is to generate regional data within each country. Hayakawa and Keola (2020), for example, used POI data in the OSM to study the recovery of 16 countries in Asia and Oceania after COVID-19.

Although, so far, none of the regional data in the IDE-GSM have been compiled using OSM data, this prospect is promising, nonetheless. First, the OSM data are updated on a daily basis. In other words, the quality of data will almost certainly improve in the future. Second, the freedom to access raw data means that it is, in principle, possible to reclassify data in a way that fits with industrial categories in the IDE-GSM. Third, combining OSM data with ground true land use data would also be likely to increase the precision of machine learning discussed in the previous section. In fact, the land use data in the OSM itself can be used as an input for machine learning.

## 4. Conclusion

This chapter traces the development of regional data compilation in the IDE-GSM. Regional data compilation started with manual inputs from ground-based surveys and censuses, before evolving to partly employ remote-sensing data and then machine-learning techniques. The application of OSM data has been explored, but it is yet to be used to generate regional data. None of these techniques is absolutely superior now or in the foreseeable future. Various combinations of the aforementioned approaches on a case-by-case basis are likely to continue to be workable solutions.

In the long run, unless the ways macroeconomic public data service is provided change fundamentally, regional data compilation in the IDE-GSM is likely to remain difficult. The following are strongly called for: The first is the spread of public data services at regional levels. Currently, macroeconomic data are often compiled at the national level, except for the European Union and some large countries. The progress of regional data development in these countries is obviously the understanding of the importance of regional data in policy analyses. A cross-country approach similar to the regional data initiative at Eurostat is strongly recommended for ASEAN. The second is to preserve and make available original public data at the regional level. In many countries, including small developing countries, regional data are collected and used to generate data at higher administrative levels. However, it is often the case that these data are discarded during the process. Preserving collected regional data would greatly benefit academic and policy-oriented studies requiring them.

Lastly, concerted efforts to promote an understanding of the importance of regional data are critically important. The availability of regional data in Europe, for example, lies in the long and strong tradition of economic geography. Economic activity occurs at

specific places, and national boundaries are often too broad to capture reality in a sensible way. It is absolutely beneficial to spread this also among developing countries.

## References

- Chen, X., and W. D. Nordhaus (2011) “Using luminosity data as a proxy for economic statistics.” *Proceedings of the National Academy of Sciences* 108(21), 8589–8594.
- Doll, C. NH, J-P. Muller, and J. G. Morley (2006) “Mapping regional economic activity from night-time light satellite imagery”, *Ecological Economics* 57(1), 75–92.
- Ebener, S., C. Murray, A. Tandon, and C. C. Elvidge (2005) “From wealth to health: modelling the distribution of income per capita at the sub-national level using night-time light imagery”, *International Journal of health geographics* 4(1), 1–17.
- Elvidge, C. D., K. E. Baugh, E. A. Kihn, H. W. Kroehl, E. R. Davis, and C. W. Davis. (1997) “Relation between satellite observed visible-near infrared emissions, population, economic activity and electric power consumption”, *International Journal of Remote Sensing* 18(6),1373–1379.
- Henderson, J. V., A. Storeygard, and D. N. Weil (2012) “Measuring economic growth from outer space”, *American Economic Review* 102(2), 994–1028.
- Keola, S., M. Andersson, and O. Hall (2015) “Monitoring economic development from space: using nighttime light and land cover data to measure economic growth”, *World Development* 66, 322–334.
- Kumagai, S., T. Gokan, I. Isono, and S. Keola (2008) “Geographical Simulation Model for ERIA: Predicting the long-run effects of infrastructure development projects in East Asia”, in Kumar, N. (ed.), *International infrastructure development in East Asia - Towards balanced regional development and integration*, ERIA Research Project Report 2007-2, IDE-JETRO, 360–393.
- Kumagai, S., K. Hayakawa, I. Isono, S. Keola, and K. Tsubota (2013) “Geographical simulation analysis for logistics enhancement in Asia”, *Economic Modelling* 34, 145–153.
- Kumagai, S., S. Keola, and T. Kudo (2012) “Policy review on Myanmar economy”, IDE-JETRO Policy Review Series on Myanmar Economy 5.
- Miyazaki, H. (2020) “A feasibility study on satellite data utilization to estimate economic indicators”, A report submitted to Institute of Developing Economies, Asian Institute of Technology, Thailand.

Tanaka, K. and S. Keola (2017) “Shedding light on the shadow economy: A nighttime light approach”, *The Journal of Development Studies* 53, no. 1, 32–48.