**Open data for global multimodal land use classification - outcome of the 2017 IEEE GRSS data fusion contest**

本文总结了2017 Data Fusion Contest（DFC17）上各参与者针对LCZ分类问题的成果。

LCZ通常应用于较粗的空间尺度（分辨率100m、200m），以便能获取城市的空间结构，而这种城市结构很难在像素分辨率很高的情况下获得。

They are generally applied at a coarse spatial scale (typically grids of resolution 100 or 200 m), in order to be able to catch this sense of urban structure that cannot be perceived when working at single-pixel scale at very high resolution.

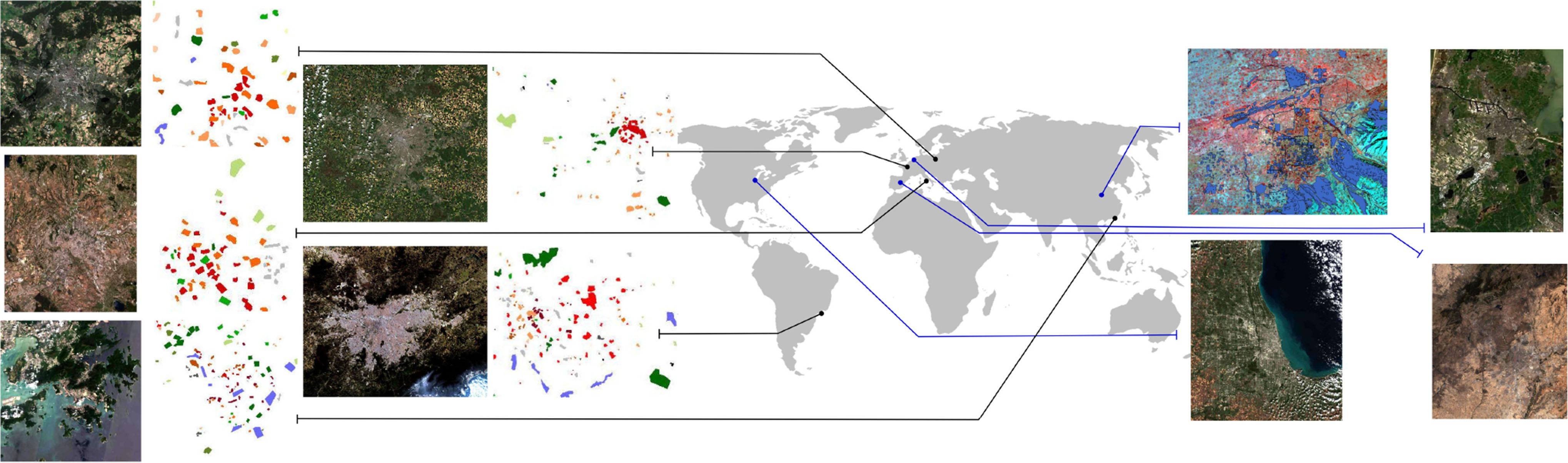
由无任何经验的志愿者提供的LCZ训练样本可能会产生精确度不够高的分类结果。

Moreover, there is evidence that training samples from unexperienced volunteers can result in inaccurate classification results.

**包含数据：**Landsat8多光谱影像、Sentinel-2多光谱影像及OSM数据。

**训练样本城市：**Berlin, Hong Kong, Paris, Rome, and Sao Paulo

**测试样本城市：**Amsterdam, Chicago, Madrid, and Xi’an



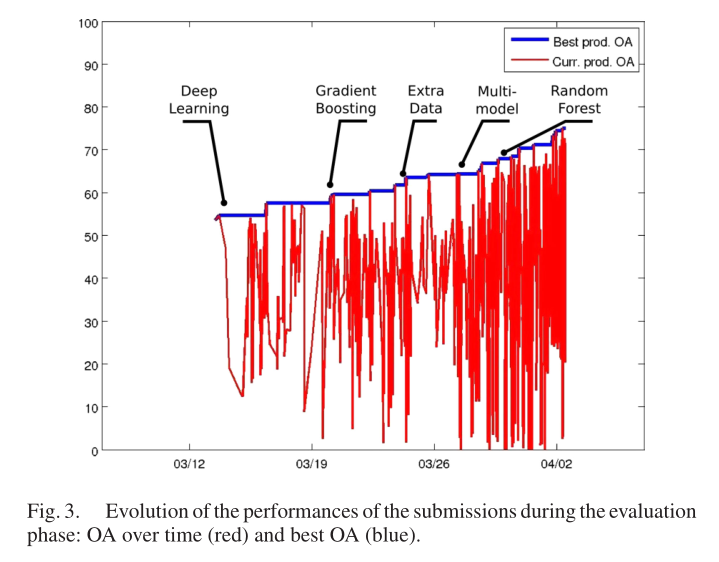
**Problems：**训练数据数量非常不均衡，其中各类数量从323到17716变化不等。

**几种类型的分类方法：**

1. Random Forest type methods
2. Boosting
3. Deep learning
4. Expert handcrafted features

**结果：**





其中采用随机森林方法的几组总体精度较高，而DL方法结果并不理想。

**原因可能是：**

实际上DFC17的数据集中训练样本的数量不少，但也并不十分地多，这可能就是为什么纯粹的DL方法所得结果的排名不高 （DL方法在面对海量训练数据时可以体现出其独特优势）。

Indeed, the number of training samples in the dataset of DFC17 was not small but not extremely large either, which was probably the reason why no pure DL approaches ranked among the winners.

从Fig. 3看出，最开始提交的DL方法在开始时划定了总体精度51.4%的基准线，结果尚可接受；但第一周之后，优势地位被Extreme gradient boost及RFs这些方法占据，结果表明，一旦调整好超参数，这些方法便能够充分利用数据的不均衡性及离散性。

It shows that DL models were ready to use soon after the release of the test data (first submissions were received less than 12h after opening the server) thus establishing an acceptable baseline with 51.4% of OA. After the first week however, the lead was taken on by teams exploiting ensemble methods like extreme gradient boost or RFs, showing that these approaches can take full advantage of imbalanced and sparse data once the right hyperparameters have been found by tuning.