Remote sensing data analysis via machine learning for land use estimation in the Greater Thessaloniki Area, Greece

Paraskevas Katsalis[1], Evangelos Bagkis[1] and Kostas Karatzas[1]

Abstract: Remote sensing data have been employed for monitoring the differences in land use over time. This information serves as the basis of any further land-related analysis, modelling and decision making. It requires satellite coverage of an area of interest, in various bands, and intense analysis of the data to correctly identify the different land types and associate them to the geographical reality precisely. In this paper, we collect Sentinel 2, level 1C satellite data to extract spectral indices and utilise them as features for land cover classification. The method is based on the use of machine learning for properly mapping the Greater Thessaloniki Area, engaging the random forest algorithm. Two different classification configurations in terms of target labels are tested for their accuracy. The main goal of the study is to present a pipeline for researchers and practitioners that need to define non-generic classes and classify geographical areas accordingly. Results, evaluated with the confusion matrix, suggest excellent performance on the test set and bring to surface limitations of the approach concerning the lack of proper high-quality data for algorithm training.

Keywords: remote sensing, satellite data, land use, machine learning, Normalized Vegetation Index

Addresses Sustainable Development Goal 13: Climate action

1. Introduction

Accurate and timely land cover (LC) information helps with land use (LU) management and LU monitoring. Identifying vegetative areas can also help detect potential landslides or forest fires [KC22]. Moreover, LC maps can help other monitoring and modelling activities that have dependence on land utilization. For example, air quality modelling is a very complex task given that sources and sinks are transient and depend partly on industrial and human activities and thus, LC data can help relevant models to associate a specific LC type with increased or decreased air quality [Jo22]. In recent years remote sensing has been a key information source for LC identification due to fine spatial resolution and great geographical coverage. The most common applications of remote sensing products aim at the identification of the changes in LU over time, mapping plant health and floods, and forest fire detection [EC17]. Furthermore, there is a plethora of spectral indices that can be derived by combining different optical bands from Sentinel 2 such as the normalized difference vegetation index [XS17]. LC and LU identification is necessary for any spatially oriented environmental management and decision-making

[1] Aristotle University, Environmental Informatics Research Group, School of Mechanical Engineering, Thessaloniki, Greece, 54124, kkara@auth.gr https://orcid.org/0000-0002-1033-5985

process. While the former maps the different types of land types of an area under study (forest, wetland, low vegetation etc.), the latter focuses on the way that various types of land are being utilized by people (i.e. a low vegetation area that is being used for agriculture or as grazing land). It is therefore essential to properly map LC as it is the basis of any further land-related analysis, modelling and decision making in scientific areas like spatial planning, microclimate zone definition, urban heat island identification, urban air quality modelling and others.

LC maps are nowadays derived from remote sensing data with a few meters resolution. Products such as Corine land cover (CLC) and Copernicus global land service are produced every six years and annually respectively. Nowadays, both products are mainly based on Sentinel 2 spectral imagery with spatial resolution close to 10 m. However, Sentinel 2 has a revisit time of approximately 5 days and thus, offers the opportunity to extract LC maps in a much finer temporal scale. For example, localization of construction sites, state of the vegetation, snow coverage and others can be beneficial in air quality modelling as they represent processes related with the production and the deposition of air pollutants [VL18, No11].

In [Ru17], the authors used multi-date spectral images and focused on classifying the vegatation with the support vector machine algorithm. Sampling training data from mutiple dates improved the robustnes of the algorithm. Similar to our study, the authors of [TV20] developed a random forest model for Thessaloniki but they concentrated on the build-up areas inside the metropolitan area. In conjuction, we focus on classifying the greater area of Thessaloniki including water bodies and agricalture areas. The authors of [Wa22] review and compare machine learning (ML) approaches such as convolutional neural networks, self-organizing maps with more traditional techniques such as cellular automata. Furthermore, [Jo19] performed a comprehensive comparison of a variety of deep learning (multilater perceptron, variational autoencoders) as well as ensemble algorithms and concluded that ensembles are more verstile and accurate than deep learning algorithms. [Hu20] concentrated in integrating remote sensing with socioeconomic data for improved land use classification. [CGR19] demostrated that convolutional neural networks can outperform traditional ML algorithms and tree-based ensembles but with increased upfront computational cost. Thus, a clear consensus of the best modeling algorithm cannot be reached as it is evident that the volume and quality of data drastically affect the performance of the models. Therefore, we choose the RF as the modeling algorithm, as it is simple to use, can run in parallel if needed and in many cases outperforms even the most advanced deep learning approaches.

Our end goal is to develop a method that will complement air quality modelling systems and provide with an updated LC map with every revisit of the satellite. We concentrated in employing already defined spectral indices related with air pollution production and deposition mechanisms. Specifically, we calculate the spectral indices associated with soil, vegetation, build-up areas, water and moisture and employ them as features to classify the study area into eight classes with the RF algorithm.

2. Materials and methods

2.1 Study area

Central Macedonia region is populated by 1.564 million people accounting to 15% of the total population of Greece, while Thessaloniki is the biggest city in the region and the second largest of Greece [Of22]. A unique characteristic of the study area is that it has a variety of LC classes consisting of industrial, urban, rural, forest, sea, lake, rivers, and agricultural fields as shown in Fig. 1 (Greater Thessaloniki Area-GTA).

Fig. 1: Area of study (using cropped Stentinel 2 Image) : the Greater Thessalobiki Area

2.2 Satellite Data

Remote sensing data were collected from the Sentinel 2 product. Sentinel 2 is equipped with 12 optical detectors that provide data in 13 bands, covering the visible, near infrared and short-wave infrared part of the spectrum. Its spatial resolution varies between 10 m, 20 m and 60 m, depending on the spectral band. This is an Earth observation satellite developed and operated by the European Space Agency (ESA). Each satellite pass is orthorectified by ESA to fit the same coordinates and is matched pixel by pixel. The geographical area of interest is identified with the aid of Sentinel's tiling grid and identified that the GTA corresponds to tile 34TFl [SE22]. This area is then cropped using a custom shapefile for better analysis of the smaller area of interest. Total dimensions of the study area are 27370 x 38730 meters.

The main and only instrument of the Sentinel 2 satellites is a multispectral instrument that generates optical images [MS22]. Using an external sensor, the assembly estimates the attitude and pointing reference to insure a 20 m pointing accuracy and then the image is taken using an optical configuration based on a Three-Mirror Anastigmat (TMA) telecentric telescope [Ca17].

There are two Sentinel 2 products that can facilitate the modelling. For this application, the Level 1C product was selected. The measurements correspond to the top of the atmosphere conditions and thus, atmospheric correction is avoided. On the other hand, Level 2A data, depict the optical properties of the Earth's surface. Both products have been tested and provided similar results. SNAP (Sentinel Application Platform) is a toolkit created by the European Space Agency for users to analyse and project the data from the Sentinel program satellites. It is a GIS application with a targeted use for those missions [SN22]. SNAP was utilised to read the data, create the main shapefile for cropping the region of interest, and to identify the coordinates and "pixel" values for training.

2.3 Spectral Indices

Apart from the 13 spectral bands taken from the satellite, another 7 spectral indices were incorporated to help the algorithm distinguish between the classes of interest. Therefore, a total of 20 features defines the input of the model. Near infrared (NIR), short-wave infrared (SWIR), RED, GREEN and BLUE are involved in the calculation of the spectral indices defined below.

- **Normalized Difference Vegetation Index (NDVI):** This index quantifies the amount of vegetation in an area by comparing near-infrared (that vegetation reflects) and red light (which vegetation strongly absorbs). NDVI is calculated as follows: NDVI= (NIR-RED)/(NIR+RED) [Pe13].

- **Normalized Difference Water Index (NDWI):** This index extracts water body information using Green and Near Infrared bands. Using this method non-water bodies have low reflectance and water bodies have high. NDWI is calculated as follows: NDWI=(GREEN-NIR)/(GREEN+NIR) [Mc96].

- **Normalized Difference Built-up Index (NDBI):** This Index highlights urban areas that have a greater reflectance in the SWIR spectral range in comparison to the NIR. NDBI is calculated as follows: NDBI=(SWIR-NIR)/(SWIR+NIR) [KC19].

- **Built-Up Index (BU):** This combines NDBI and NDVI, to automatically map built-up areas resulting in a map where only built-up and barren area pixels have a positive value. BU is calculated as follows: BU=NDBI-NDVI [Ch10].

- **Bare Soil Index (BSI) :** This index is used to determine the bare soil tiles in a digital satellite image. It uses 4 spectral bands (SWIR, NIR, RED, BLUE). It is a

normalized index with higher values indicating higher chance of bare soil: BSI=[(SWIR+RED)(NIR+BLUE)]/[(SWIR+RED)+(NIR+BLUE)]. [Di17].

- **Green Chlorophyll Index (GCI):** The use of this index is for estimation of the chlorophyll across a wide range of plant species. Here it is used to help the algorithm as a second index for high vegetation areas GCI=(NIR/GREEN)-1[Gi03].

- **Moisture Stress Index (MSI):** This Index is used for canopy stress analysis using the SWIR over NIR ratio. Higher values indicate greater plant stress while lower show less moisture content. It is calculated as follows: MSI=(SWIR/NIR) [HR89].

The normalized indices and the BSI range between -1 and 1. The range for BU is [-1.307, 0.767], for GCI is [-0.832, 10.99] and for MSI is [0.196, 10].

2.4 Methodology

The input of the RF classifier was constructed by concatenating the initial 13 bands with the 7 extracted spectral indices, creating a 20-feature input vector for each "pixel". Eight major classes were identified and included as labels for the classification namely, water, shallow water, concrete, brick, dry soil and crop, wet oil, low vegetation, and tree cover. For the water and shallow water 250 locations were annotated. For concrete and brick classes 574 locations, 1374 for dry soil and crop, 1052 for wet soil and 1120 for low vegetation and 1000 for tree cover with the SNAP tool. The total number of pixels in the study area for the algorithm to classify was 10.600.401. A total number of 6194 locations were manually classified into their respective categories. To overcome the fact that not all bands are in the desired resolution, an up-sampling inter-area interpolation method was used. Thus, we obtained all the channels in the 10-meter resolution.

To move on with the modelling, we selected the RF algorithm as it is robust and accurate. RF [Br01] is an ensemble of decision trees that provides estimations with reduced variance compared to each individual tree. That's because decision trees suffer from overfitting with high variance however, when the estimations are combined with the RF algorithm this variance is reduced leading to improved estimations. Each tree is grown on a randomized subset (random feature selection and bootstrap) of the initial dataset and tries to estimate the same target. During inference, the RF combines the individual estimations via majority voting to produce the final estimation. The number of estimators was set to 100, maximum depth to 5 and all the other parameters were kept to their default values as defined in the scikit-learn python implementation of the algorithm. Finally, the classification map was saved in .tiff format using the metadata of the RED band. For the visualization a custom colormap was created to better represent the natural colours of each category for better visual representation. The performance of the model was evaluated on a total of 619 locations of the manually annotated data. More specifically, a number of 75 locations per class were selected (i.e. 75 x 8 = 700 locations), as well as with 19 additional key locations, which were selected because of the difficulty that the algorithm had in

categorizing them. The 619 locations amount to 10% of the training locations.

3. Results and discussion

In Fig. 2, three of the calculated spectral indices are depicted. With regards to the green coverage, incorporating the NDVI (panel c) and GCI (panel a) indices light up areas with chlorophyll, help to better differentiate between vegetation and non-vegetated areas. Furthermore, NDWI (panel b) offers valuable information for the model to distinguish between land and water.

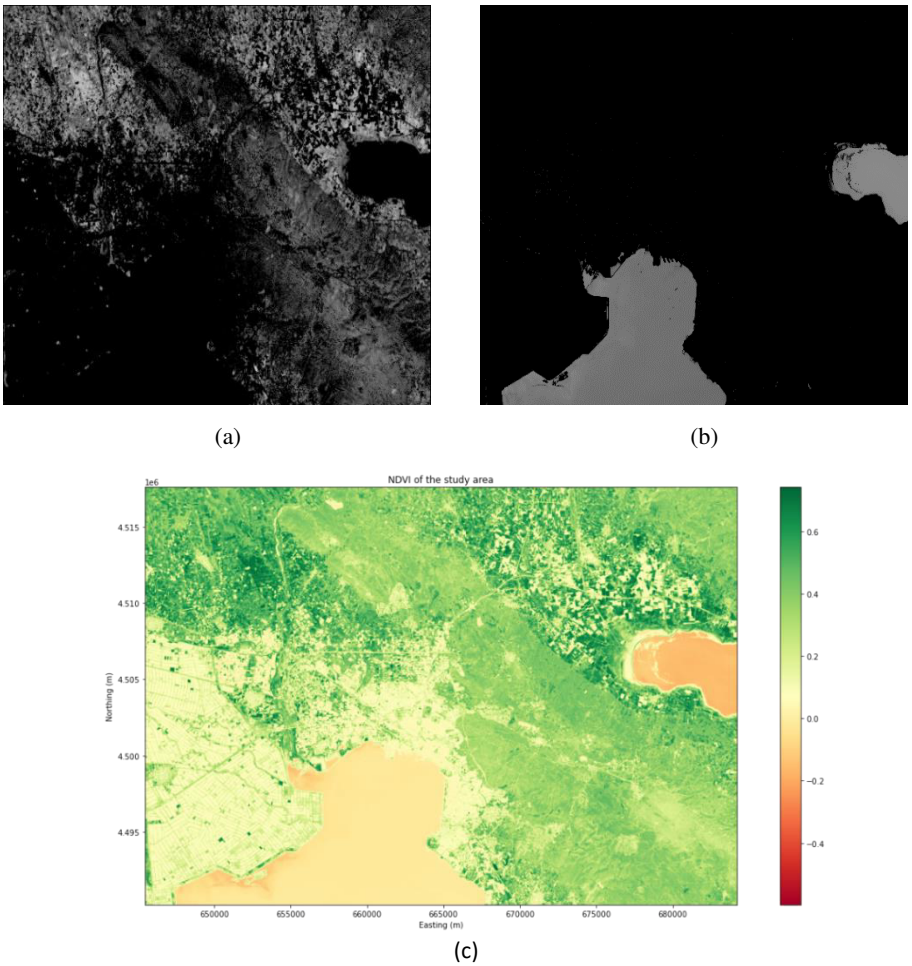

Fig. 2: Spectral indices visualization, a) GCI, b) NDWI, c) NDVI

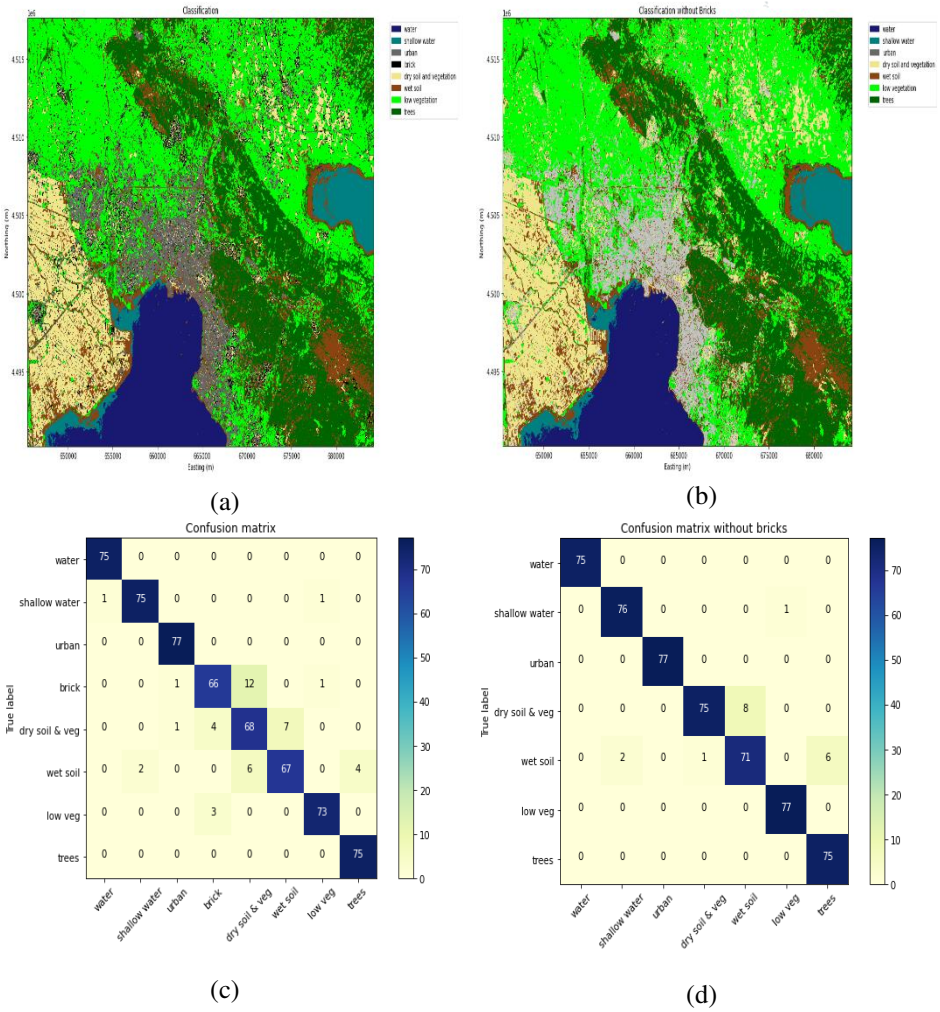

Fig. 3: RF classification results, a) depicts the land-cover map with 8 classes including the Brick class, b) shows the same map but without the Brick class, c) concfusion matrix with Brick class, d) confusion matrix without Brick class.

Overall, the accuracy, calculated on the test set, shows good agreement. However, brick, dry soil, and wet soil categories have very similar reflectance values in almost all the bands and RF struggles to differentiate between them as is evident in Fig. 3 (c). The algorithm can easily identify concrete urban areas due the high reflectance of construction materials and the geometrical shapes that makes these areas stand out. Wet soil areas can be identified around the sores, regarding the city's port on the northern part and the Koroneia lake on the eastern part. Some wet soil areas can be seen in the mountainous region and

are thought to be successfully categorized. Low vegetation fields and dry vegetation fields can be differentiated. Highways even though they have a relatively small width of one to two pixels are easily identified. In the confusion matrix it can be observed that 5 out of 8 classes have almost perfect classification with the problem being in the three classes that have similar reflectance in the spectral bands and indices. When the brick class is removed, the accuracy increases from 93% to 96% but this might lead to the misclassification of the brick covered locations. In future studies we plan to adopt a more robust validation scheme to increase the trust in the results and to compare RF with potentially better algorithms.

4. Conclusions

The main goal of this study was to present a methodology that can help researchers produce LC and LU maps every time the Sentinel 2 revisits a specific area with minimal (>400 locations as a rule of thumb) manual annotation with the aid of ready to use tools. A reproducible machine learning pipeline was presented for LC classification. We proposed to include seven spectral indices that are easily derived from Sentinel 2 L1C as features during modelling to improve the classification error. Furthermore, we identified that brick, dry soil, and wet soil have similar reflectance properties and the misclassifications occur for these classes whereas all other classes were reconstructed almost perfectly. One important aspect related to the increase in accuracy observed after removing the brick class, is that care should be given in the choice of LC classes, and we propose that the classes should be considered with the end goal in mind. It is preferable to identify the classes that play the most important role for the specific problem and build a classifier thereafter than using generic LC maps with irrelevant classes. Therefore, having a machine learning pipeline, like the one proposed here, at hand, can potentially help users create more custom maps for specific use cases. Finally, the proposed method makes use of graphical tools and is low-code, making it easy to work for practitioners in need of accurate and specific user defined classes.

Bibliography

[Br01] Brienman S.: Random Forests. Machine Learning 45, 5–32, 2001.

[CGR19] Carranza-García M.; García-Gutiérrez J.; Riquelme J.: A Framework for Evaluating Land Use and Land Cover Classification Using Convolutional Neural Networks. *Remote Sensing*, *11*(3), 274, 2019.

[Ca17] Cazaubiel V.; Chorvalli V.; Miesch C.: The multispectral instrument of the Sentinel2 program, Proc. SPIE 10566, International Conference on Space Optics — ICSO 2008, 105660H, 21 November 2017.

[Ch10] He, C.; Shi, P.; Xie, D.; Zhao, Y.: Improving the normalized difference built-up index to map urban built-up areas using a semiautomatic segmentation approach. *Remote Sensing Letters*, *1*(4), 213-221, 2010.

[Di17] Diek S.; Fornallaz F.; Schaepman ME.; De Jong R.: Barest Pixel Composite for Agricultural Areas Using Landsat Time Series. Remote Sensing. Remote Sens. 2017, 9, 1245, 2017.

[EC17] Emery W.; Camps A.: Introduction to Satellite Remote Sensing, Chapter 10 - Land Applications p. 701-766 , Elsevier, 2017.

[Gi03] Gitelson, A.; Gritz, Y.; Merzlyak, M.: Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. *Journal Of Plant Physiology*, *160*(3), 271-282, 2003.

[Hu20] Huang Z.; Qi H.; Kang C.; Su Y.; Liu Y.: An Ensemble Learning Approach for Urban Land Use Mapping Based on Remote Sensing Imagery and Social Sensing Data. *Remote Sensing*, *12*(19), 3254, 2020.

[HR89] Hunt J. E.; ROCK, B.: Detection of changes in leaf water content using Near- and Middle-Infrared reflectances. *Remote Sensing Of Environment*, *30*(1), 43-54, 1989

[Jo22] Johansson L.; Karppinen A.; Kurppa M.; Kousa A.; Niemi J.; Kukkonen J.: An operational urban air quality model ENFUSER, based on dispersion modelling and data assimilation, Environmental Modelling and Software, to appear.

[Jo19] Jozdani S.; Johnson B.; Chen D.: Comparing Deep Neural Networks, Ensemble Classifiers, and Support Vector Machine Algorithms for Object-Based Urban Land Use/Land Cover Classification. *Remote Sensing*, *11*(14), 1713, 2019.

[KC22] Kalsnes, B.; Capobianco, V.: Use of Vegetation for Landslide Risk Mitigation. *Springer Climate*, 77-85, 2022.

[KC19] Kuc, G.; Chormanski, J.: SENTINEL-2 IMAGERY FOR MAPPING AND MONITORING IMPERVIOUSNESS IN URBAN AREAS. ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. XLII-1/W2. 43-47. 10.5194/isprs-archives-XLII-1-W2-43-2019, 2019.

[Mc96] McFEETERS, S.: The use of the Normalized Difference Water Index (NDWI) in the delineation of open water features. *International Journal Of Remote Sensing*, *17*(7), 1425-1432, 1996.

[MS22] MSI instrument Overview https://sentinels.copernicus.eu/web/sentinel/technical-

guides/sentinel-2-msi/msi-instrument, last accessed 2022/5/24

[No11] Novotny E.; Bechle M.; Millet D.; Marshall J.: National Satellite-Based Land-Use Regression: NO2 in the United States , Environmental Science & Technology 2011 45(10), p. 4404-4414, 2011.

[Of22] Official General Secretariat of National Statistical Service of Greece Web site https://www.statistics.gr/el/statistics/-/publication/SAM03, last accessed 2022/5/06.

[Pe13] Pettorelli, N.: The normalized difference vegetation index, Oxford University Press, 2013.

[Ru17] Rujoiu-Mare M.; Olariu B.; Mihai B.; Nistor C.; Săvulescu I.: Land cover classification in Romanian Carpathians and Subcarpathians using multi-date Sentinel-2 remote sensing imagery, European Journal of Remote Sensing, 50:1, 496-50, 2017.

[SE22] Sentinel 2 User Guide https://sentinels.copernicus.eu/web/sentinel/user-guides/sentinel-2-msi, last accessed 2022/5/06.

[SN22] SNAP website https://step.esa.int/main/toolboxes/snap/, last accessed 2022/5/06.

[TV20] Tsolakidis I.; Vafiadis M.: Urban land cover mapping, using open satellite data. Case study of the municipality of Thessaloniki. OP Conf. Ser.: Earth Environ. Sci. 410 012062, 2020.

[VL18] Vizcaino P.; Lavalle C.: Development of European NO2 Land Use Regression Model for present and future exposure assessment: Implications for policy analysis, Environmental Pollution, Volume 240, p. 140-154, 2018.

[Wa22] Wang J.; Bretz M.; Dewan M.; Delavar M.: Machine learning in modelling land-use and land cover-change (LULCC): Current status, challenges and prospects. *Science Of The Total Environment*, *822*, 153559, 2022.

[XS17] Xue, J.; Su, B.: Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications, Journal of Sensors , vol 2017, 2017