Optimization and Engineering (2021) 22:1783–1805
https://doi.org/10.1007/s11081-021-09668-2

RESEARCH ARTICLE

Optimization of air pollution measurements with unmanned aerial vehicle low-cost sensor based on an inductive knowledge management method

Sławomir Pochwała[1] · Stanisław Anweiler[1] · Adam Deptuła[2] · Arkadiusz Gardecki[3] · Piotr Lewandowski[1] · Dawid Przysiężniuk[1]

Received: 31 December 2020 / Revised: 19 July 2021 / Accepted: 20 July 2021 /
Published online: 30 July 2021

Abstract
The article presents the study of Particulate Matter air pollution with PM_1, $PM_{2,5}$ and PM_{10} by means of a low-cost sensors mounted on Unmanned Aerial Vehicles. The article is divided into two parts. In first part pollution measurement system is described. In second part expert system for optimization of flight parameters is described. The research was conducted over a municipal cemetery area in Poland. The obtained results were analyzed through an inductive knowledge management system (decision tree method) for classification analysis of air pollution. The decision tree mechanism would be used to optimize flight parameters taking into account the air pollution parameters. The analysis was made from the influence of PM concentration point of view, depending on the altitude. The decision tree method was used, which allowed to determine, among other aspects, which PM indicator should be measured and which altitude plays a greater role in the optimization of air pollution measurements by means of cheap sensors mounted on drones. As a result of the analysis, the optimum flight altitude of the measurement drone in the specified area was determined.

Keywords Air pollution · Low-cost sensor · UAV · Particulate matter · Decision tree · Inductive knowledge management

1 Introduction

In urban areas Particulate Matter (PM) is a key issue affecting personal pollution exposure levels (Cao et al. 2020). In Poland the tradition of burning candles and lamps at cemeteries is very cultivated and candles are burned basically all year

✉ Stanisław Anweiler
s.anweiler@po.edu.pl

Extended author information available on the last page of the article

round, and especially around Catholic holidays the number of burned candles can be measured in many millions. Ground level air quality studies using bio-indicators such as moss show a significant problem in this regard (Ciesielczuk et al. 2012). Therefore, an attempt has been made to conduct air quality studies at higher altitudes using UAVs. Air quality monitoring has traditionally been conducted using sporadically distributed, costly reference monitors (Tanzer et al. 2019). To increase the resolution of air pollution measurements the use of Unmanned Aerial Vehicles, as a platform for air pollution surveys, can be applied for a wide range of research scenarios (Chilinski et al. 2016). Especially light weight gas sensor systems are suitable for airborne applications (Ahlawat et al. 2019). Current developments in miniaturization of chemical equipment and in low-cost small drones are catalyzing exponential growth in the use of such platforms for environmental chemical sensing applications (Javier and Marc 2020; Johnson et al. 2020). The evolution of low-cost sensors (LCSs), whose prices currently range from a few to several EUR, has made the spatio-temporal mapping of the indoor and outdoor air pollution possible but the diversity of them for various applications make their optimum selection challenging (Omidvarborna et al. 2021). Generally, the approaches to investigating the fine-scale spatiotemporal distribution of air pollutants can be classified into three categories: air pollution dispersion models, fixed-site measurements, and mobile measurements (Cao et al. 2020). Tracing of atmospheric pollutants release of and validation of the measurements using unmanned aerial vehicles is still difficult task from a mathematical point of view (Šmídl and Hofman 2013; Yungaicela-Naula et al. 2019; Villa et al. 2016).

With the rapid development of a new branch of distributed measurement using low-cost sensors and multiple UAV platforms, there has been a huge increase in air quality data. In order to reduce the amount of data and increase the significance of measurements, it will be necessary to introduce decision-making rules into this process. Increasing amount of data requires management optimization. Various methods for information classification and decision support are available in the literature (Deptuła and Partyka 2017; Beshah Tesema et al. 2005). The theory of approximate sets with high data classification efficiency methods is presented in the paper (Nowicki et al. 1992a, b). One helpful tool for determining strategy is a decision tree. The decision tree mechanism can be used to optimize the flight parameters by taking into account the environmental data. It is not the operator who should think at what height and at what speed should he perform the measurement flight but the software itself should suggest optimal flight parameters in order to minimize the influence of the sensor movement on the measured values. Decision tree theory is one of the basic methods of inductive knowledge acquisition and exploration, which can be used for classification and prediction process. The method allows to determine the association of a given object to homogeneous classes in relation to the dependent variable on the basis of measurements of one or more predictor variables. As a consequence, it is possible to classify objects on the basis of which the tree was built, as well as to use the created classification rules for subsequent predictions. In the case of applied decision trees, knowledge acquisition is based on the analysis of samples, with each such sample being described by a set of attributes on the basis of which classification rules are built. The novelty here is the approach to air pollution measurements during motion. In the beginning of the research the impact of environmental data has not been clearly characterized. In this paper an attempt to apply known and generally accepted methods, as a tool selection for gathering data optimization has been done.

In recent years, researchers have been exploring the interdependencies between different areas of knowledge. With this interdisciplinary approach, a certain point has been reached that requires the integration and systematization of accumulated knowledge. The evolution of management and design support systems has resulted in the application of this knowledge to solutions based on artificial intelligence and machine learning. Such efforts have allowed the automation of management accuracy assessment based on design principles using knowledge transformation, and consequently to the emergence of Intelligent Decision Systems and Expert Systems. All the time the decision making process must use the latest available information technologies. In the decision support systems for design and management methodology the decision-making process has a specification of the framework process, which is often simplified and modified, depending on the importance and subject matter of the decision or design task. By building an appropriate decision model, even random problems can be solved and the results obtained are satisfactory in most cases of technical problems.

There is a wide range of research into the development of methodologies supporting decision-making processes and management control, design methodologies and systems of varying scale of complexity including artificial intelligence. Statistical classifiers as applied to UAVs are a very broad concept and can address areas such as image recognition in various spaces (Pi et al. 2020), also applied in logistics (Raj and Sah 2019), building engineering (Omar and Nehdi 2017), agriculture (Moysiadis et al. 2021), and detection of physical and chemical particles. A special direction of development that strengthens the role of classification systems is the combination of various processing methods, inference and seeking knowledge developed separately under artificial intelligence into one coherent hybrid consulting system. In particular, there are two general approaches to creating hybrid decision support systems: CI—Computational Intelligence and SC—Soft Computing (Liu et al. 2010).

At the same time, the state of the art in machine learning is constantly increasing. It covers the problems of constructing systems whose operation increases with the experience represented by the set of teaching examples. In this area, particular attention should be paid to the methods of trees and graphs (Rutkowski et al. 2012; Liu et al. 2016; Pijls and Bruin 2001; Iordanov 2010). In the technical problems under consideration, one of the methods of classification of information and decision support is the method of inductive rule generation using decision trees. In the case under review, entropy is used to determine the most relevant attribute. Inductive decision trees can be compared in the process of classification, prediction and determining the importance of decision variables with multivalued logical trees. There are many works presenting the use of inductive decision trees, multi-valued logical trees and multi valued logical equations as decision support tools in discrete optimization and determining decision variables (Deptuła and Partyka 2017,2011; Deptula and Partyka 2018). Classification methods found wide application in the analysis and diagnostics of internal combustion, electric and hybrid engines. For example, in

Wu and Liu (2009), a system for diagnosing damage to internal combustion engines using wavelet packet transformation (WPT) and artificial neural network (ANN) techniques has been proposed (Staszewski et al. 1997; Specht 1991). Currently, most of the research focuses on data mining algorithms, although other stages are equally important for the successful implementation of the whole process (Correia et al. 2003; Jayamalini and Ponnavaikko 2017; Chen 2015; Linoff and Berry 2011). There are many different descriptions in the literature of both individual steps and entire design and decision support processes (Pijls and Bruin 2001; Horzyk 2012). Inductive decision trees play an important role in practical applications. Inductive decision trees can be compared with multivalued logic trees in the process of classification, prediction and rank determination of decision variables. Such studies exist, but so far they have not been applied to the analysis of drone measurements. In this case, common objective functions include the so-called Gini index (Dixon et al. 1987). Which can be helpful in hierarchizing quantities for air quality measurements using UAVs.

The main objective of the conducted research was to develop a decision model based on inductive classification rules. A drone flight data record was analyzed, then knowledge was extracted from the measurements and finally induction trees were generated. The induction tree was then used to generate decision rules. These steps were used to build an expert system that would ultimately make the flight altitude decisions. Eventually, the system itself should make these decisions. Although at this stage of research, the system is designed to assist the drone pilot in making decisions about optimal flight parameters.. Unlike classical computer programs, the knowledge contained in the tree describes the problem domain without providing a detailed way to solve the problem (algorithm). This is important for the research that was analyzed in this paper. Finally, the generated induction trees made it possible to create and save a knowledge base. Then, an expert system was built based on this knowledge base to inform the user what action to take.

2 Materials and methods

In the following sections of the paper, two issues are raised. The first is a description of an unmanned aircraft system with a low-cost sensor to measure air quality. The second is a characterization of the expert system applied later to optimize the flight parameters of the unmanned system. The combination of these two issues is a novelty and the authors' contribution to the development of the scientific field.

A study was carried out on the concentration of PM1, PM2,5 and PM10 particles over the municipal burial site of the deceased in a medium-sized suburban town in Poland. Figure 1 shows the measurement site. An attempt was made to determine the influence of this specific holiday on the air quality, when families of the deceased visit the graves of their loved ones en masse and light candles to their memory. The huge number of candles lit introduces a great number of chemical compounds into the atmosphere. This location was chosen because of the considerable distance from busy roads and the lack of influence of additional sources of air pollution, which could affect the results of the measurements.

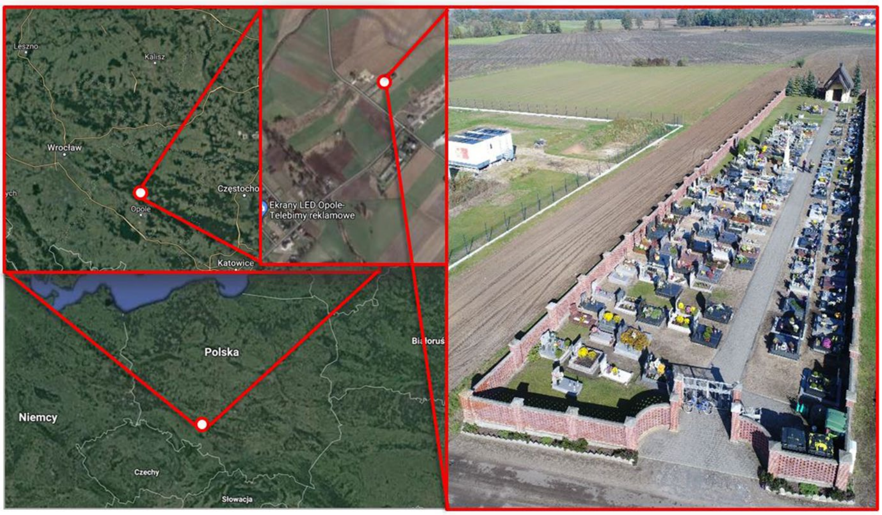


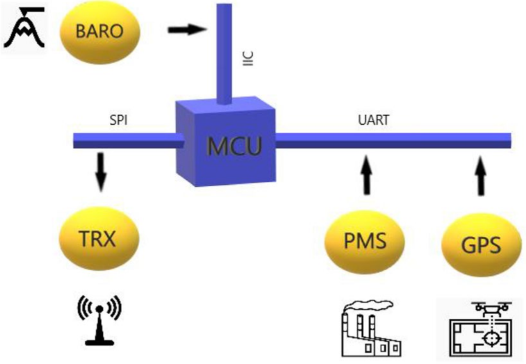
Fig. 1 Location of the object on which air quality measurements were made

2.1 Description of the UAV system

Figure 2 shows a diagram of the air quality measurement system necessary to achieve the assumed objectives for use on unmanned aerial vehicle (drone). Figure 2 shows a simplified scheme of the device and is split into three basic modules: measurement, data and receiver. The measurement and data module are attached to the UAV and the receiver enables to receive data on-line.

In Fig. 2 the BARO refers to the baromether, UART, IIC and SPI are data buses connected to the MCU – Microcontroller Unit, TRX refers to wireless connection to the data receiver, PMS refers to PMS7003 sensor (for measurements of PM1, PM2.5, PM10) which was used for air sample acquisition, and GPS is the

Fig. 2 Low-cost sensor system diagram



global positioning system sensor. Table 1 provides detailed information on the sensor used in low-cost multisensor systems.

When the device measures the concentration of particulate matter (PM), it can operate up to 12 h on battery power. The sensors are linked to on-board voltage dividers, which allocate the voltage to a 10-bit analog-to-digital converter. This is provided by built-in ATmega 328 microcontroller (MCU). The spatial position of the sensing array is acquired from the built-in NMEA protocol provided by the GPS receiver mounted on the universal asynchronous receiver-relay (UART) port and processed by the MCU. The geo-localization can also be obtained by the drone's onboard GPS system, but the sensor's built-in system allows to make 3D maps directly from the sensor readings. Sampling frequency is 1 Hz; every second the location and pollution data are saved on a secure digital card (SD card) in comma separated values file format (CSV file). At the same time the data is transferred in the ultra high frequency (UHF) band (433 MHz). The transmitter uses simple on/off (OOK) modulation. A small power of about 50 mW is sufficient to provide a stable connection in line of sight (LOS) flight mode. The data rate of the STX882 transmitter is between 0.1 and 9.6 kbps. The data rate has been set to 1.8 kbps because a lower speed has better anti-interference parameters and is sufficient to hold all data at an interval of one second. The transmitter acts as a bridge between the computer/smartphone and the sensor module. It receives data from the system on the drone and retransmits this data via Bluetooth. This is useful when plotting data on the screen in real time. It is also possible to connect the receiver module directly to the computer using a USB cable. The total cost of the developed device is below 30 €. Figure 4 shows the original design of the printed circuit board (PCB) used for the low-cost installation of the PM sensor and the measurements that are presented in this work. This is the next prototype. The first PCB prototypes were presented in Pochwała et al. (2020). The PCB has a very simple design and is based mainly on external modules such as external GPS module. Individual modules of the system are as follows: MCU;GPS; barometric altimeter; voltage converter and regulator circuit; FSK transmitter—wireless data transmission—does not write to SD; PMS dust sensor adapter; connectors. The board used in the research described in this paper was an improved version presented in Pochwała et al. (2020). The improvement concerned primarily the layout and, consequently, further reduction of dimensions. As a consequence, smaller and cheaper UAVs can be used to carry the sensor. For the presented prototypes the whole software of the ATmega328 microcontroller was written in Arduino IDE using the programming language C++.

The designed mass of the constructed device is limited, because the drone can only lift a specified load. The sensors are packed into a small chamber with forced

Table 1 Summary of basic technical data of the low-cost PM sensor used

Sensing device	Pollution type	Range of measurement	Data output	Power consumption	Resolution	Cost
PMS 7003	$PM_{1.0}$, $PM_{2.5}$, PM_{10}	0–1000 µg/m^3	Digital	< 500 mW	1 µg/m^3	13 €

air circulation through an internal fan. The sensor housing is an open structure, whose elements are made in 3D printing technology. The mass of the measuring system was 78 g and of the separate power supply system 100 g. The total mass of the device was 178 g and did not exceed the maximum UAV starting mass. In comparison with the previous prototype of the system, the mass was reduced by 22%.

Flights were carried out with the DJI Matrice 200 drone. This UAV is designed for industrial applications, can fly at wind speeds up to 10 m/s and at temperatures down to − 20 °C and allows to fly up to 38 min at speeds up to 83 km/h. The mass of the used drone varies between 3.8 and 4.5 kg and can lift from 1.6 up to 2.3 kg of cargo, depending on the battery used. The drone's construction meets the IP43 standard. It is able to fly in very harsh environmental conditions, such as heavy rain or strong dust. It has sensors installed at the front, top and bottom, which allows to plan autonomous flights. During the flight it uses GPS and GLONASS networks.

The measurements were taken in the following order: 2 autonomous flight plans were made using Litchi software. The first route had 9 waypoints with a total length of 358 m. A flight was made on the designated route at five altitudes. The flight altitude was h = 20, 25, 30, 40, 50 m, respectively. The second route consisted of 8 waypoints with a total length of 420 m and the same altitudes as the first route. Figure 3 shows the respective routes of autonomous flights.

2.2 Application of decision-making rules

Generally, decision rules are induced from data sets representing information about a set of objects called learning examples, described by a set of attributes. Most algorithms look for such rules through inductive generalization of the description of the examples of learners—see discussions in Beshah Tesema et al. (2005), Swe and Sett (2019), Estivill-Castro and Murray (1998).

A decision tree is a structure which has ordinary properties of trees in the meaning assigned to the tree in the information technology, so it is a structure composed of nodes from which branches come to other nodes or leaves. It is convenient to

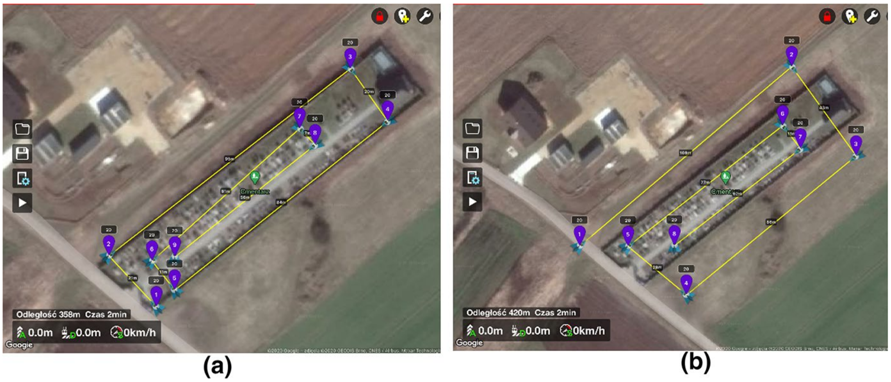


Fig. 3 A sample query "query-by-example" to the decision-making system

define tree structures in a recursive way. Assuming that a given branch X on which attributes a_1, a_2,…, a_n and the set of notions in the category C are determined:

1. The leaf containing any category label $d \in C$ is a decision tree.
2. If t: $X \rightarrow R_t$ is a test data made from the example values of attributes with a set of possible results $R_t = \{ r_1, r_2, \dots r_m \}$, then the node containing the test t, from which m branches come out, given that for $i = = 1, 2, \dots, m$ branch i corresponds to the result r_i and leads to the tree T_i, is a decision tree.
3. For any node of n decision tree t_n is a test data connected with it, and for each of its possible results $r \in R_t$ by $n_{[r]}$ node or child leaf, to which the n branch related to the r result leads from the node. The notation described above is presented in Fig. 4

Information included in the set of training examples is equal to (see also Fig. 9):

$$I(E) = -\sum_{i=1}^{|E|} \frac{|E_i|}{|E|} \log_2 \left(\frac{|E_i|}{|E|} \right) \tag{1}$$

where E–the set of training examples, $|E_i|$–the number of examples which describe i object, $|E|$–the number of examples in the training set E.

The expected value of information after the division of the set of examples E into subsets $E^{(m)}, m = 1, \dots, |V_a|$, for which the attribute a has the value V_m, determined as Quinlan (1986) and Mitchell (2006):

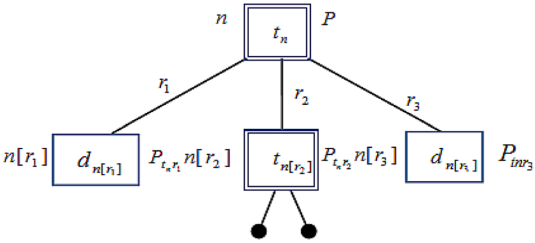
$$I(E, a) = \sum_{m=1,K,|V_a|,E^{(m)}=\emptyset} \frac{|E^{(m)}|}{|E|} I\left(E^{(m)}\right) \tag{2}$$

where $|E^{(m)}|$—the number of examples after the division of the set E in relation to the value m of a given attribute, $|E|$—the number of examples in the training set E.

The decision tree was used to generate the knowledge base. The expert system created using *PCShell* tool consists of the following components shown in Fig. 5

At this stage of research, the system is designed to assist the drone pilot in making decisions about optimal flight parameters like altitude and velocity. In this case, the optimal flight parameters are GPS coordinates as shown in the resulting section of the paper. The complete program is responsible for: controlling source extraction and activation, data acquisition and preprocessing, access

Fig. 4 View of the autonomous flight plans: first route (**a**); second route (**b**)



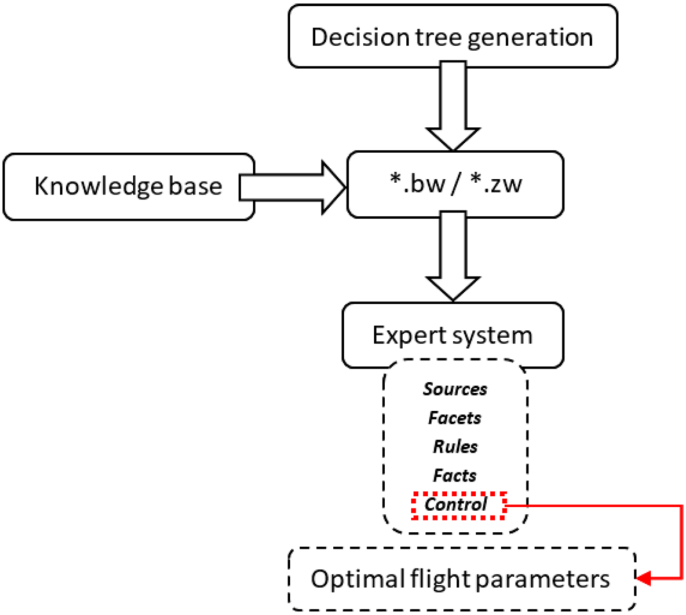


Fig. 5 Horizontal PM2.5 concentration profile

to database files, for dynamic data exchange, etc. However, the most important part of the program is the Control block, which provides the user with the optimal parameters for which he should perform measurement flights. Figure 6 shows the detailed diagram of the developed decision support system with an expert program.

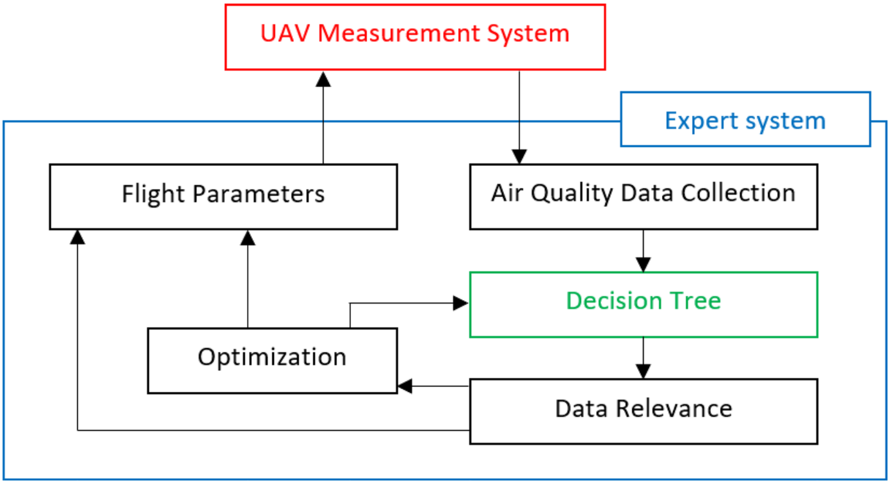


Fig. 6 General block diagram of the developed system

The entire expert system consists of three main components. The first one marked with a red frame is related to the UAV system and measurement data collection. The second one marked with a green frame concerns the decision tree. The third one, surrounded by a blue frame, connects both of the above elements and creates a complete expert system informing the user about the optimal flight parameters. The expert system, using the measurement data and decision tree methodology through feedback, is constantly evolving increasing its efficiency through continuous improvement of the decision making quality. From the knowledge engineering point of view, it is important that the dynamic change of certain knowledge base parameters can be done automatically, without direct user intervention in the knowledge base source. The use of an inductive decision tree allows for quick classification of examples and maximum characterization descriptions, giving a full description of the phenomenon.

Because this type of expert systems can be classified according to many different factors, such as the design, the method of deduction or the type of data processed. In addition, expert systems are also divided into deterministic, which process certain knowledge, and non-deterministic or probabilistic, which operate on approximate, uncertain knowledge. For these reasons, it is difficult to compare each expert system with each other. However, a brief comparison of the advantages and disadvantages of the developed original expert system based on *PCShell* tool with the widely used *CLIPS* system, written in the *C* programming language at NASA laboratories in 1984, can be found in Table 2 (CLIPS 2021).

3 Results and discussion

As a result of the conducted tests, air pollution with particulate matter in a specific area was measured and air pollution profiles were visualized for specific altitudes. The sensor was calibrated. The calibration was carried out on the basis of comparing the minimum and maximum readings from the reference pollution measurement station with the measurement results from the developed sensor. The calibration station is a part of the Polish national environmental monitoring system, whose measurement data are commonly available. The results of calibration measurements are shown in Fig. 7.

Table 2 Basic properties characterizing the developed *PCShell*-based expert system and the *CLIPS*-based system

Properties of the system	*PCShell*	*CLIPS*
Requirement to know the internal skeleton language	No	Yes
The requirement for consistency when creating an interface	No	Yes
Keeping consistency in data representation	No	Yes
Minimum size of the knowledge base	No limits	50 records
Easy data entry via built-in editor	Yes	No
Built-in deduction mechanism	Yes	No

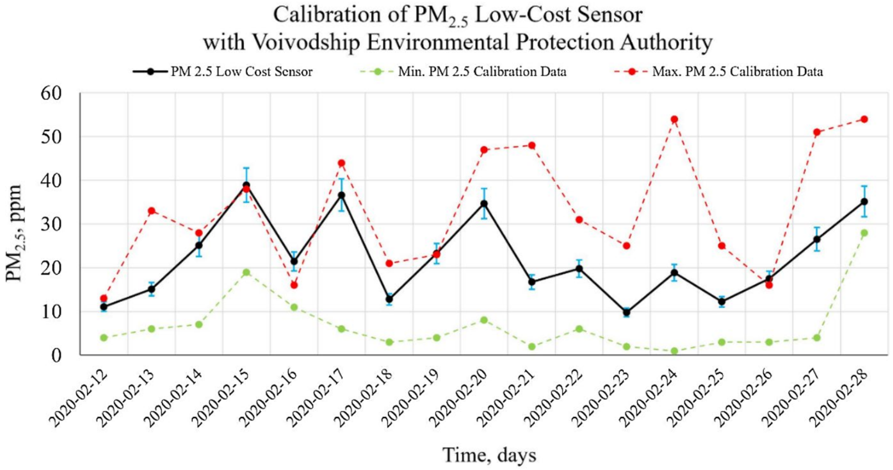


Fig. 7 Detailed scheme of the decision support system with the expert program

Figure 7 shows comparative graphs of measurements made with the developed low-cost system against the background of measurements from the reference meteorological station. It can be seen that in the examined two-week period the results from the cheap sensor are within the scattering of readings from accredited sensors in the reference weather and air pollution measurement station. Sampling time for the low-cost setting is 15 s, while the reference station gives 15-min averages. So, to show the trend for the reference station the measurement data were averaged. We did this to show the trend because in some cases low-cost sensors may tend to overestimate the measured values. This averaging is very important to check if the low-cost laser sensor does not get dirty over time, which can lead to a measurement deterioration. More accurate calibration analyses are in line with other authors (Nguyen 2019).

Measurements using a cheap UAV mounted particulate matter sensor were made on October 28, 2020. Figures 8, 9, 10 show the concentration of contaminants over the studied area in the form of horizontal concentration profiles for PM1, PM2.5 and PM10 at altitudes of 20, 25, 30, 40 and 50 m respectively, made according to the routes shown in Fig. 3.

In the presented drawings you can clearly see the layered nature of the profile. Additionally, it is possible to notice a shift in the concentration of pollutants according to the direction of the wind, as shown, for example, in Fig. 11.

Next the use of inductive decision-making systems in the classification of the state of pollution was made. The set of analyzed data (learning files) is represented in the form of a decision board as a record in first order logic. Formally, an information board is a pair according to Eq. (3).

$$IT = (U, A) \tag{3}$$

where U is a non-empty and finite set of objects, A is a non-empty and finite set of attributes. The V_a set is an attribute domain. In our case, the attributes of the

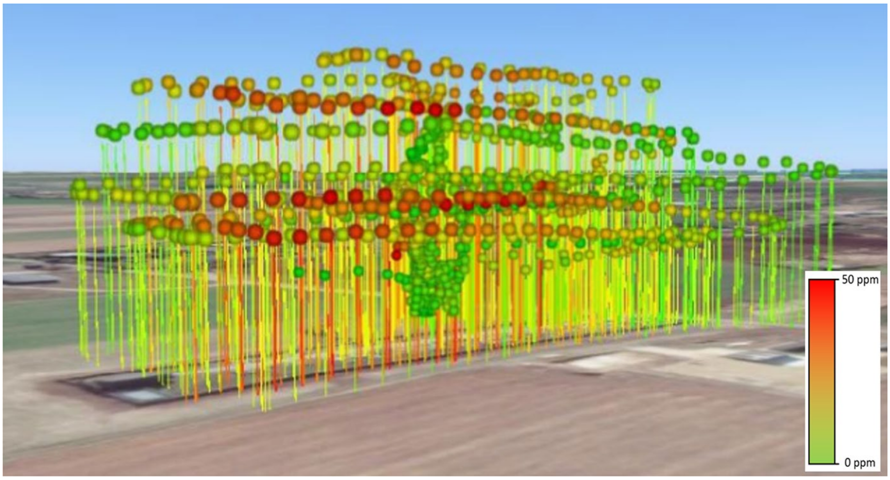


Fig. 8 Calibration data from a two-week measurement period. Comparison of minimum (green) and maximum (blue) readings from the reference pollution measurement station with the measurement results using the developed sensor (black)

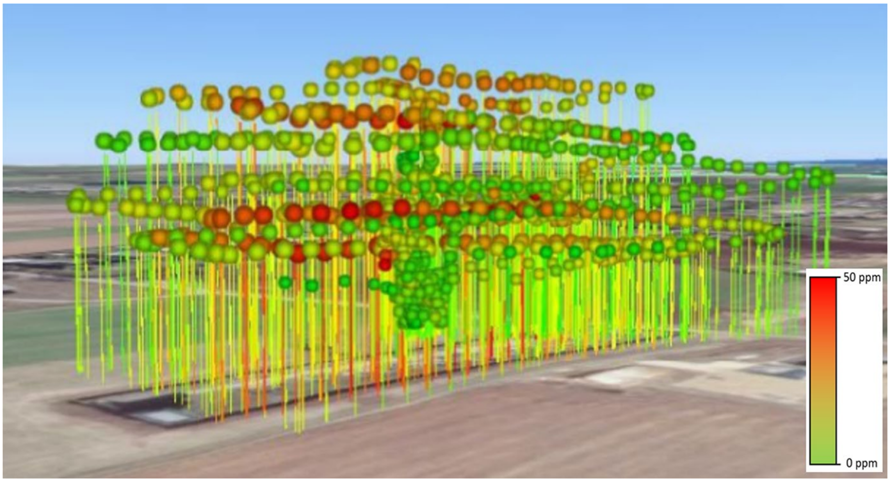


Fig. 9 Horizontal PM1 concentration profile

numerical type have been adopted, whose domains are defined on numerical measurement scales: interval or quotient (on these scales, as opposed to the ordinal scales, arithmetic operations are also possible).

A separate category are the attributes whose domains are arranged according to preferences, i.e. they are criteria. The analysis was made from the two criteria point of view:

I. Analysis of the influence of measurement data concentration (*PM1*, *PM2,5* and *PM10*) depending on the position at a given altitude.

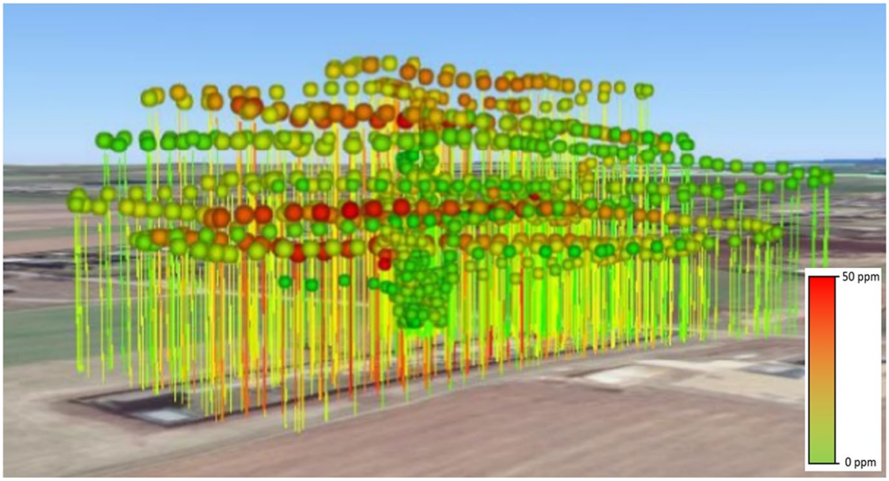
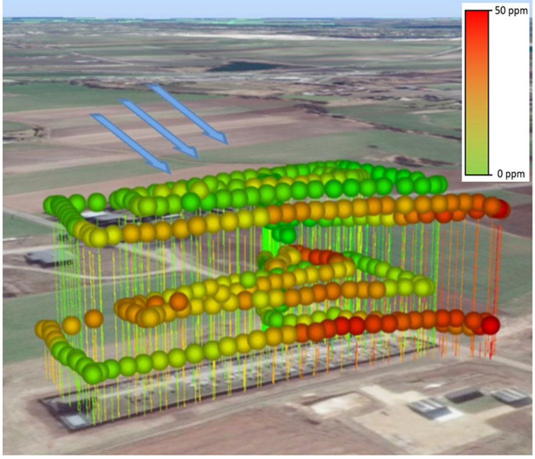


Fig. 10 Horizontal PM10 concentration profile

Fig. 11 Example of wind influence on PM10 concentration distribution



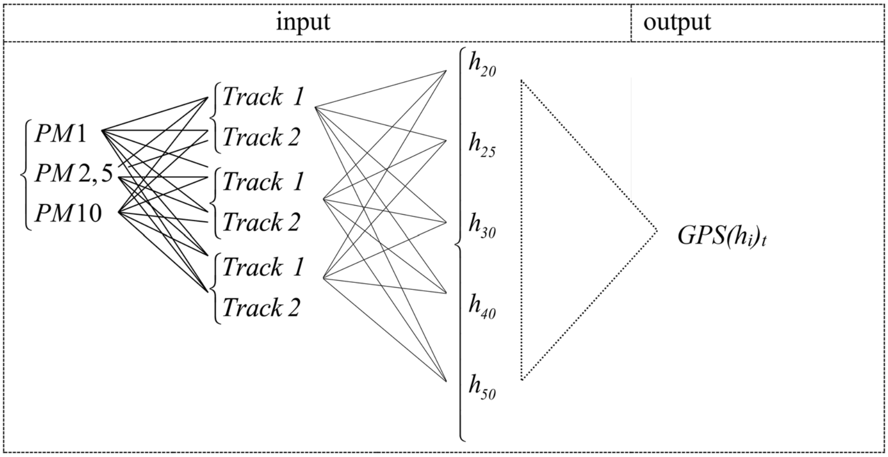
II. Analysis of GPS altitude and specific PM measurements (i.e. what is the most important altitude and GPS measurement for a given *PM1*, *PM2,5* and *PM10* for both routes).

Analysis I: The analysis was made for two routes: Track_1 Track_2, for each route 5 heights were taken into account: Track_1: $h_1=20$, $h_2=25$, $h_3=30$, $h_4=40$, $h_5=50$ and Track_2: $h_1=20$, $h_2=25$, $h_3=30$, $h_4=40$, $h_5=50$.

In the analysis we assume that: (1) ***input*** attributes are measurements of *PM1, PM2,5, PM10* for altitude data h_1, h_2, h_3, h_4, h_5 for both *Track 1* and *Track 2*; (2) the ***output*** attributes are the $GPS(h_i)_t$ reading values for the given *hi* altitude and for the given *t* route (Table 3).

For the analysis *DeTreex* software was used, which enables induction of decision trees. In order to prepare the data, the teaching (.lrn) and testing (.tst) files were

Table 3 Structural model of analysis I



built. A total of 10 data files were received during the measurement (Track 1—for h_{20}, h_{25}, h_{30}, h_{40}, h_{50}, Track 2—for h_{20}, h_{25}, h_{30}, h_{40}, h_{50}).

Based on the above, appropriate learning tables were obtained for all combinations. For tables that are learning files, decision rules and then induction trees were generated in the first step. The problem of finding a minimum set of rules that covers the set of examples and classifies them correctly is NP- complete. Evidence uses the transformation of this problem to the problem of minimum rulebook coverage (Andersen and Martinez 1995; Stefanowski and Vanderpooten 2001). In Mitchell (1999) a general scheme of such algorithms are given.

In the next stage, induction trees are generated. Figure 12 shows an example of an inductive decision tree for Track 1(h_{20}).

The other input trees for Track 1(h25), Track 1(h30), Track 1(h40), Track 1(h50), Track 2 (h20), Track 2 (h25), Track 2 (h30), Track 2 (h40), Track 2 (h50) are not included in the article due to their volume. The results of their analysis are discussed below.

The conclusions of the inductive decision making systems concern two routes. For the first route (Track 1) the value measured PM1 plays the most important role. Therefore, because it occurs in the inductive roots of the tree at the height of 20 m, 40 m, and 50 m.

PM10 plays a further role because it occurs in the roots of trees at heights 25 m and 30 m. Additionally, inductive decision trees showed the most important values of these measurements. This means for:

- PM1- for height 20 m—value 11, for height 40 m—value 10, for height 50 m—value 11.
- PM10- for height 25 m—value 17, for height 50 m—value 17.

In the meaning of the hierarchical classification means that these values played the most important role in determining the values of pollutants at given altitudes on

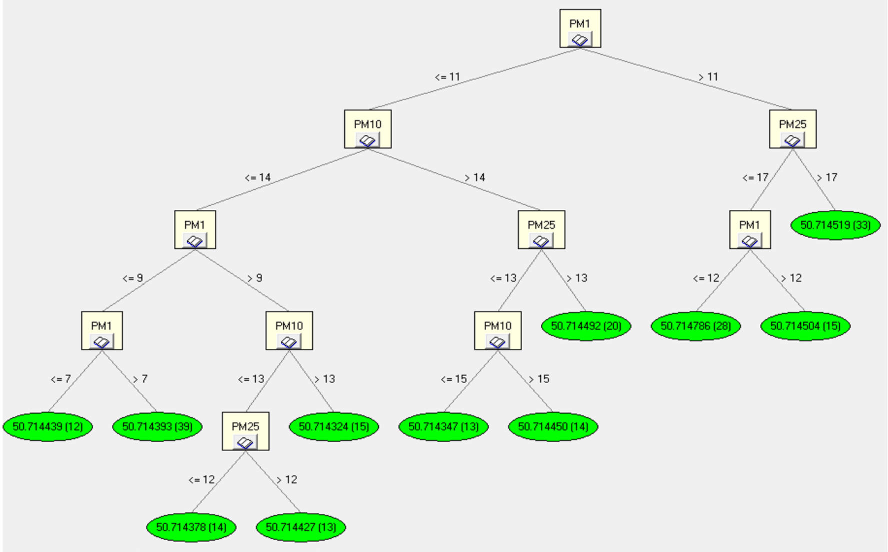


Fig. 12 Inductive decision tree for Track 1(h_{20})

the route (Track 1). In addition, the optimal GPS measurement values are given in the leaves of induction trees. This means the right place on the route.

For the second route (Track 2), the PM2,5 measured value plays the most important role this time. The first route was not taken into account at all in the classification analysis. It occurs in the roots of induction trees at heights 30 m, 40 m and 50 m. For the other heights there is PM1. Additionally, inductive decision trees showed the most important values of these measurements. This means for

- PM25 for height 30 m—value 14, for height 40 m—value 18, for height 50 m—value 18.
- For PM10 for height 25 m—value 13, for height 50 m—value 13.

As with route 1, the GPS values for route 2 (Track 2) are listed.

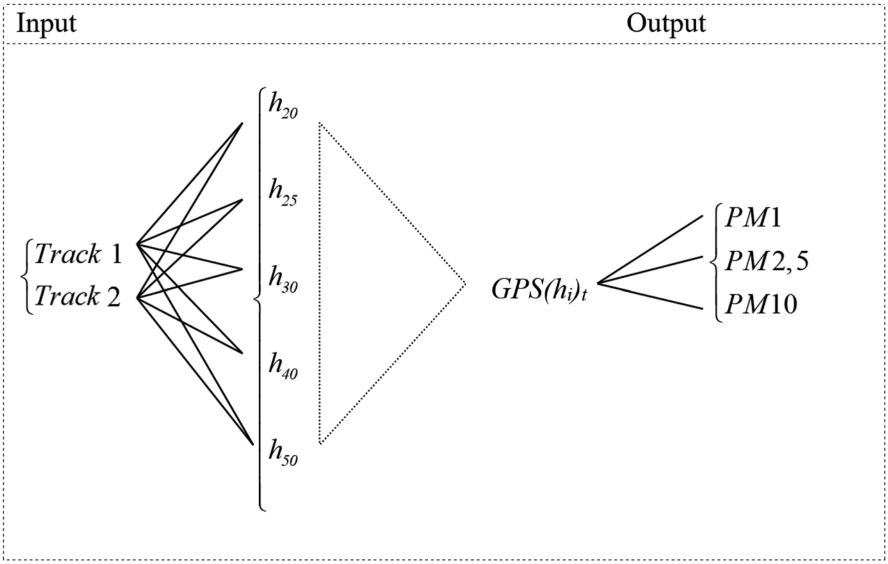
Analysis II was made also for two routes: Track_1 Track_2, for each route 5 heights were taken into account: Track_1: h1 = 20 m, h2 = 25 m, h3 = 30 m, h4 = 40 m, h5 = 50 m and Track_2: h1 = 20 m, h2 = 25 m, h3 = 30 m, h4 = 40 m, h5 = 50 m.

In the analysis we assume that (1) ***input*** attributes are two types of Track 1 and Track 2, altitudes h1, h2, h3, h4, h5, and GPS(hi)t reading values; (2) ***output*** attributes are values of PM1, PM2,5, PM10 for height data h1, h2, h3, h4, h5 for both routes (Table 4).

A total of 6 data files were received (Track 1—PM1, PM2,5, PM10 and for Track 2—PM1, PM2,5, PM10).

As in analysis I, for each case we have obtained the tables of learning files, database file and induction tree.

Table 4 Structural model of analysis II



The conclusions of the inductive decision making systems concern two routes. In case of the analysis answering the question how the height influences the values of PM1, PM2,5 and PM10 contamination, the parts of the induction tree (twigs) which are minimal on the given tree should be considered. The most important height value for the measurements is then generated.

For route 1 (Track 1), the optimal height for PM1 is 40 m and 50 m. (Fig. 13, side shortest twigs on the tree).

In the leaves, on the other hand, optimal values of these measurements are generated (for PM1 to 9 and 10).

They analyse the remaining induction trees, then for Track 1

- The optimum height for PM2.5 is also 40 m and 50 m (Fig. 14) with optimum values of 12,
- For PM10 the optimum height is 25 m and 30 m (Fig. 15) with values 13 and 15.

For Track 2, the optimal height for

- PM1 is 20 m, 40 m and 50 m with values 14, 19, 14 (Fig. 16). This influence is not so visible here,
- PM2,5 is 40 m and 50 m and even more 50 m (Fig. 17),
- PM10 is 30 m (Fig. 18).

An additional conclusion for analysis II is that for route I (Track 1), it is easier to establish a "clear" relationship between altitudes h20, h25, h30, h40, h50 and values PM1, PM2,5 and PM10 than for route 2 (Track 2).

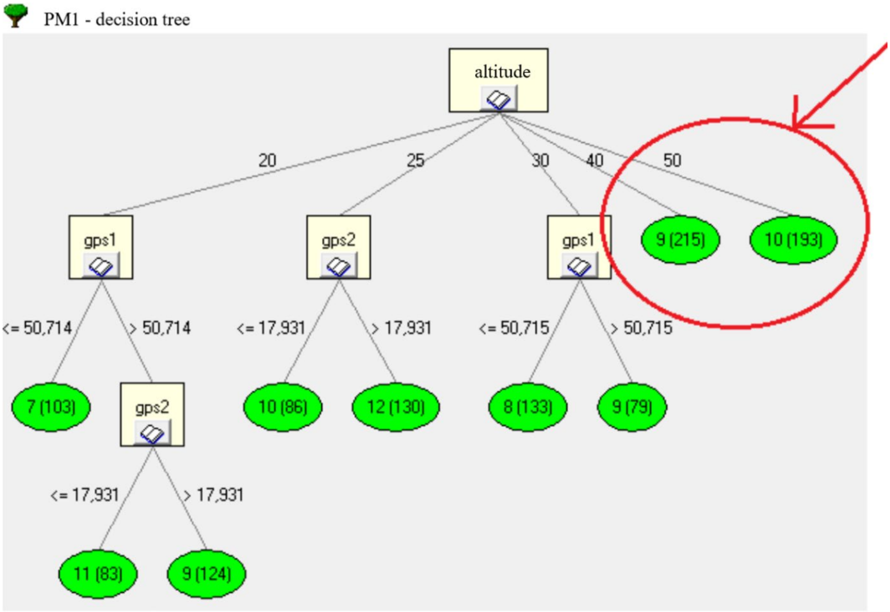
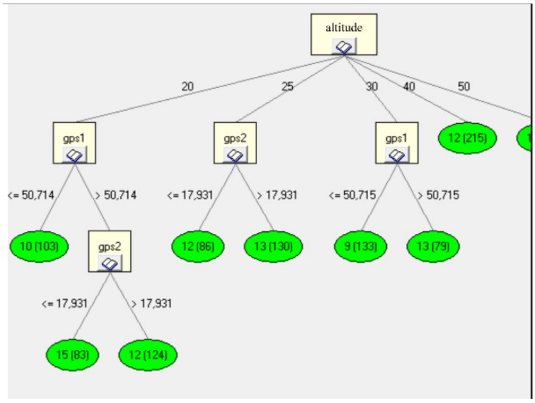


Fig. 13 Inductive decision tree for Track 1 with designated shortest branches of the tree indicating the optimal altitude

Fig. 14 A fragment of the learning file (table) for Track 1(PM2,5)



From the trees, an expert system is created. So, in the authors' opinion, there is no need to consider additional computations at this stage. When generating inductive trees, recursive computations are used at each split until the maximum depth is reached. The choice of the maximum depth is very important and significantly affects how the model may underfit the data.

Figure 19 shows the data fit for the tree shown in Fig. 12 (Fig. 12—Inductive decision tree for Track 1(h20)).

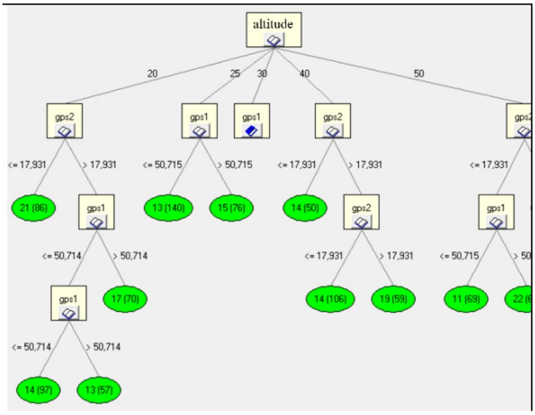


Fig. 15 A fragment of the learning file (table) for Track 1(PM10)

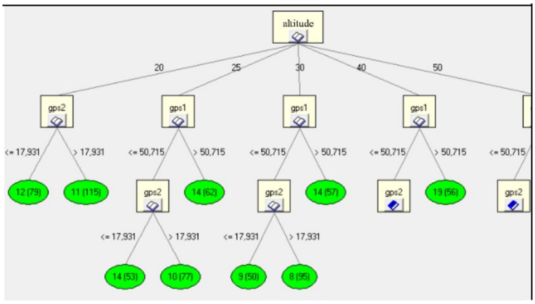


Fig. 16 A fragment of the learning file (table) for Track 2(PM10)

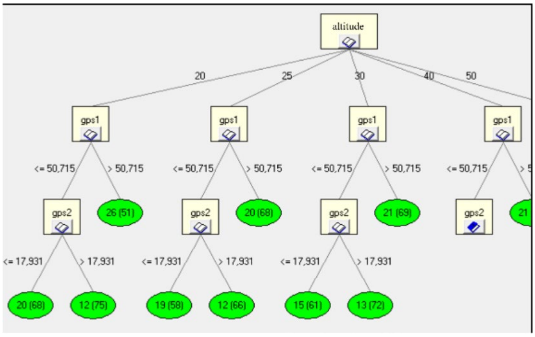


Fig. 17 A fragment of the learning file (table) for Track 2(PM2,5)

The graphs in Fig. 19 show the translation of classification trees into regression tree notation. For regression trees, typical objective functions include mean square error, mean absolute error, or standard deviation. When classifying (learning) a tree, a function is needed in the program. To represent the performance of the tree in another way, the standard deviation reduction function can be used to build a regression model. This function gives results equivalent to the mean square error. Then, instead of entropy, these graphs are used to create a cost

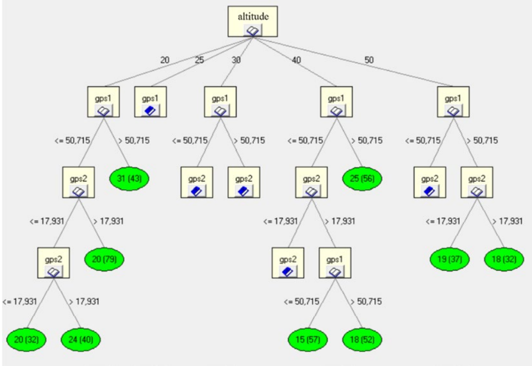


Fig. 18 A fragment of the learning file (table) for Track 2(PM10)

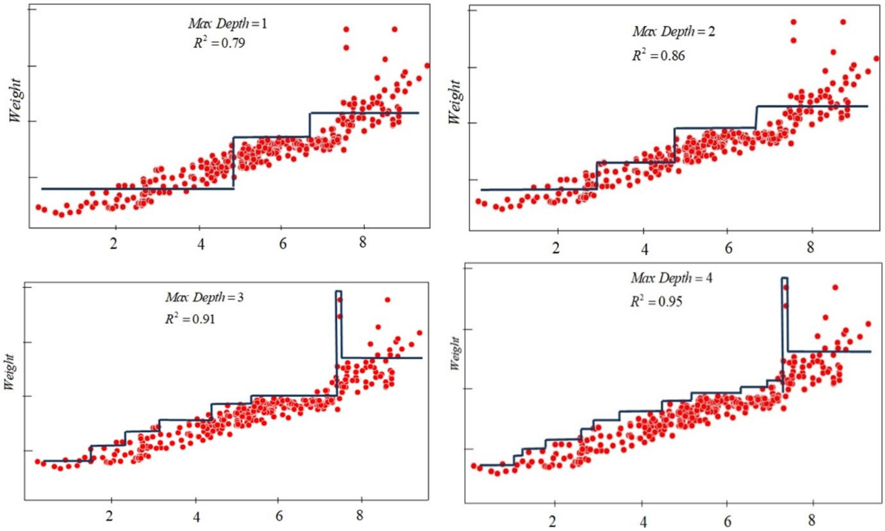


Fig. 19 Effects of changing the maximum tree depth

function that is minimized when learning the expert system. The decision tree algorithm used to perform the splits is called the classification and regression tree algorithm.

This is applied recursively at each entry to the next level of the tree until the maximum depth is reached. Our tree model will be recursively invoked multiple times from each tree model produced. Each node in the tree performs a binary division (unless it is a leaf), which means that each node creates two divisions (left and right). Each split creates a new decision tree, and each new tree performs successive left and right splits. This process repeats until the decision tree is deep enough.

In summary, the graphs in Fig. 19 show the accuracy of classification with testing depth. The tree previously shown in Fig. 12 (Fig. 12 Inductive decision tree for

Track 1(h20)) ended up fitting at R = 0.95. The red dots in the graphs are the measured data. Depth is the search step.

Finally, after applying the proposed expert system, the number of measurement points (GPS coordinates) is limited to the most relevant ones. This determines the optimal route on which to take the measurements to get the same results which also reduces the execution time. This is shown in Fig. 20.

4 Conclusions

The air pollution spreads unevenly across the earth's atmosphere depending on the many factors, such as height and season of the year and time of the day, weather conditions (wind in particular). The horizontal profile of pollutant distribution should be monitored due to the possibility of accumulation in different layers of the atmosphere.

Developed air pollution multisensory arrangement is an easy to create and user-friendly, low-cost device (approx. 30 €) with high measurement potential. The investigation has shown that inexpensive sensors can be a good basis of information on the atmospheric condition at a range of altitudes. This statement has been proven by paralleling readings from designed sensor setup with the accredited instruments used by governmental air pollution monitoring system. It has also been confirmed that low-cost sensors are unaffected by contamination and can operate over extended periods without maintenance. Created small and lightweight, cheap, devices can be successfully used for continuous and repeated measurements by creating atmosphere quality inspections, especially in cities and heavily urbanized regions. It is not difficult to imagine a swarm of such devices performing coordinated examination flights to check the condition of the atmosphere. This is far more significant than depending on one or two ground-based measuring points, which are currently used in cities

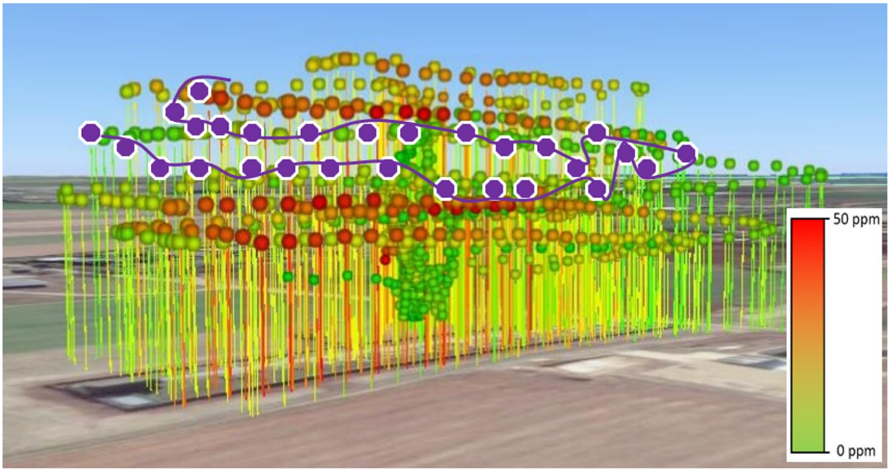


Fig. 20 Optimized PM measurement route (violet color)

in Poland. The final result of the optimization of the environmental data collection with the use of low-cost sensor UAVs was the determination of the measurement flight altitude. For particles with larger sizes (PM10), the optimum measurement flight altitude is 30 m. For particles with smaller sizes (PM1 and PM2.5) the optimal sampling altitude is the range of 40–50 m. These results are confirmed for all test series.

References

Ahlawat A, Mishra SK, Gumber S, Goel V, Sharma C, Wiedensohler A (2019) Performance evaluation of light weight gas sensor system suitable for airborne applications against co-location gas analysers over Delhi. Sci Total Environ 697:134016. https://doi.org/10.1016/j.scitotenv.2019.134016

Andersen T, Martinez T (1995) Learning and generalization with bounded order rule sets

Beshah Tesema T, Abraham A, Grosan C (2005) Data mining using adaptive regression trees I. J Simul 6:11

Cao R, Li B, Wang Z, Peng ZR, Tao S, Lou S (2020) Using a distributed air sensor network to investigate the spatiotemporal patterns of PM2.5 concentrations. Environ Pollut. https://doi.org/10.1016/j.envpol.2020.114549

Chen G (2015) Application of web data mining technique to enterprise management of electronic commerce. In: Proceedings—2014 7th international symposium on computational intelligence and design, ISCID 2014, vol 1. pp 154–157. https://doi.org/10.1109/ISCID.2014.103.

Chilinski MT, Markowicz KM, Markowicz J (2016) Observation of vertical variability of black carbon concentration in lower troposphere on campaigns in Poland. Atmos Environ 137:155–170. https://doi.org/10.1016/j.atmosenv.2016.04.020

Ciesielczuk T, Olszowski T, Prokop M, Kłos A (2012) Application of mosses to identification of emission sources of polycyclic aromatic hydrocarbons. Ecol Chem Eng S 19(4):585–595. https://doi.org/10.2478/V10216-011-0041-8

CLIPS (2021) A tool for building expert systems. http://www.clipsrules.net/. Accessed 16 July 2021

Correia JH, Wille R, Stumme G, Wille U (2003) Conceptual knowledge discovery-a human-centered approach. Appl Artif Intell 17(3):281–302. https://doi.org/10.1080/713827122

Da Wu J, Liu CH (2009) An expert system for fault diagnosis in internal combustion engines using wavelet packet transform and neural network. Expert Syst Appl 36(3):4278–4286. https://doi.org/10.1016/j.eswa.2008.03.008

Deptuła A, Partyka MA (2011) Application of dependence graphs and game trees for decision decomposition for machine systems. J Autom Mob Robot Intell Syst 5(4):17–26

Deptuła A, Partyka MA (2017) Inductive decision tree analysis of the validity rank of construction parameters of innovative gear pump after tooth root undercutting. Int J Appl Mech Eng 22(1):25–34. https://doi.org/10.1515/ijame-2017-0002

Deptula A, Partyka MA (2018) application of complex game-tree structures for the HSU graph in the analysis of automatic transmission gearboxes. J Mach Eng 18(4):96–113. https://doi.org/10.5604/01.3001.0012.7713

Dixon PM, Weiner J, Mitchell-Olds T, Woodley R (1987) Bootstrapping the Gini coefficient of inequality. Ecology 68(5):1548–1551. https://doi.org/10.2307/1939238

Estivill-Castro V, Murray AT (1998) Discovering associations in spatial data-an efficient medoid based approach. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), vol 1394, pp 110–121. https://doi.org/10.1007/3-540-64383-4_10

Horzyk A (2012) Information freedom and associative artificial intelligence. In: Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 2012, vol 7267 LNAI, no. PART 1. pp 81–89. https://doi.org/10.1007/978-3-642-29347-4_10.

Iordanov B (2010) HyperGraphDB: a generalized graph database. In Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics), 2010, vol 6185 LNCS. pp 25–36. https://doi.org/10.1007/978-3-642-16720-1_3.

Javier B, Marc S (2020) Environmental chemical sensing using small drones: a review. Sci Total Environ. https://doi.org/10.1016/J.SCITOTENV.2020.141172

Jayamalini K, Ponnavaikko M (2017) Research on web data mining concepts, techniques and applications. In: 2017 international conference on algorithms, methodology, models and applications in emerging technologies, ICAMMAET 2017, vol 2017-Janua. pp 1–5. https://doi.org/10.1109/ICAMMAET.2017.8186676.

Johnson BJ, Malanoski AP, Erickson JS (2020) Development of a colorimetric sensor for autonomous, networked, real-time application. Sensors 20(20):1–21. https://doi.org/10.3390/s20205857

Linoff G, Berry MJA (2011) Data mining techniques : for marketing, sales, and customer relationship management. Wiley

Liu S, Duffy AHB, Whitfield RI, Boyle IM (2010) Integration of decision support systems to improve decision support performance. Knowl Inf Syst 22(3):261–286. https://doi.org/10.1007/s10115-009-0192-4

Liu Q, Li Y, Duan H, Liu Y, Qin Z (2016) Knowledge graph construction techniques. Comput Res Dev 53(3):582–600. https://doi.org/10.7544/issn1000-1239.2016.20148228

Mitchell TM (1999) Machine learning and data mining. Commun ACM 42(11):30–36. https://doi.org/10.1145/319382.319388

Mitchell T (2006) The discipline of machine learning. Pittsburgh, PA 15213: Carnegie Mellon University, CMU-ML-06–108

Moysiadis V, Sarigiannidis P, Vitsas V, Khelifi A (2021) Smart farming in Europe. Comput Sci Rev 39:100345. https://doi.org/10.1016/j.cosrev.2020.100345

Nguyen B (2019) PM2. 5 low-cost sensors and calibration data for SDS011. https://doi.org/10.13140/RG.2.2.12945.68966.

Nowicki R, Słowiński R, Stefanowski J (1992a) Evaluation of vibroacoustic diagnostic symptoms by means of the rough sets theory. Comput Ind. https://doi.org/10.1016/0166-3615(92)90048-R

Nowicki R, Słowiński R, Stefanowski J (1992b) Rough sets analysis of diagnostic capacity of vibroacoustic symptoms. Comput Math with Appl. https://doi.org/10.1016/0898-1221(92)90159-F

Omar T, Nehdi ML (2017) Remote sensing of concrete bridge decks using unmanned aerial vehicle infrared thermography. Autom Constr 83:360–371. https://doi.org/10.1016/j.autcon.2017.06.024

Omidvarborna H, Kumar P, Hayward J, Gupta M, Nascimento EGS (2021) Low-cost air quality sensing towards smart homes. Atmosphere (basel) 12(4):453. https://doi.org/10.3390/atmos12040453

Pi Y, Nath ND, Behzadan AH (2020) Convolutional neural networks for object detection in aerial imagery for disaster response and recovery. Adv Eng Informatics 43:101009. https://doi.org/10.1016/j.aei.2019.101009

Pijls W, De Bruin A (2001) Game tree algorithms and solution trees. Theor Comput Sci 252(1–2):197–215. https://doi.org/10.1016/S0304-3975(00)00082-7

Pochwała S, Gardecki A, Lewandowski P, Somogyi V, Anweiler S (2020) Developing of low-cost air pollution sensor—measurements with the unmanned aerial vehicles in Poland. Sensors 20(12):3582. https://doi.org/10.3390/s20123582

Quinlan JR (1986) Induction of decision trees. Mach Learn 1(1):81–106. https://doi.org/10.1007/bf00116251

Raj A, Sah B (2019) Analyzing critical success factors for implementation of drones in the logistics sector using grey-DEMATEL based approach. Comput Ind Eng 138:106118. https://doi.org/10.1016/j.cie.2019.106118

Rutkowski L, Korytkowski M, Scherer R, Tadeusiewicz R, Zadeh LA, Zurada JM (eds) (2012) Artificial intelligence and soft computing, vol 7267. Springer, Berlin

Šmídl V, Hofman R (2013) Tracking of atmospheric release of pollution using unmanned aerial vehicles. Atmos Environ 67:425–436. https://doi.org/10.1016/j.atmosenv.2012.10.054

Specht DF (1991) A general regression neural network. IEEE Trans Neural Netw 2(6):568–576. https://doi.org/10.1109/72.97934

Staszewski WJ, Worden K, Tomlinson GR (1997) Time-frequency analysis in gearbox fault detection using the Wigner-Ville distribution and pattern recognition. Mech Syst Signal Process 11(5):673–692. https://doi.org/10.1006/mssp.1997.0102

Stefanowski J, Vanderpooten D (2001) Induction of decision rules in classification and discovery-oriented perspectives. Int J Intell Syst 16(1):13–27. https://doi.org/10.1002/1098-111X(200101)16:1%3c13::AID-INT3%3e3.0.CO;2-M

Swe SM, Sett KM (2019) Knowledge discovery in classification and distribution of butterfly species from Dagon University Campus, Myanmar by Rule Induction: CN2 algorithm. Int J Trend Sci Res Dev 5:600–603. https://doi.org/10.31142/ijtsrd26380

Tanzer R, Malings C, Hauryliuk A, Subramanian R, Presto AA (2019) Demonstration of a low-cost multi-pollutant network to quantify intra-urban spatial variations in air pollutant source impacts and to evaluate environmental justice. Int J Environ Res Public Health. https://doi.org/10.3390/ijerph16142523

Villa T, Salimi F, Morton K, Morawska L, Gonzalez F (2016) Development and validation of a UAV based system for air pollution measurements. Sensors 16(12):2202. https://doi.org/10.3390/s16122202

Yungaicela-Naula N, Garza-Castañon LE, Zhang Y, Minchala-Avila LI (2019) UAV-based air pollutant source localization using combined metaheuristic and probabilistic methods. Appl Sci 9(18):3712. https://doi.org/10.3390/app9183712

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Sławomir Pochwała[1] · Stanisław Anweiler[1] · Adam Deptuła[2] · Arkadiusz Gardecki[3] · Piotr Lewandowski[1] · Dawid Przysiężniuk[1]

Sławomir Pochwała
s.pochwala@po.edu.pl

Adam Deptuła
a.deptula@po.edu.pl

Arkadiusz Gardecki
a.gardecki@po.edu.pl

[1] Faculty of Mechanical Engineering, Opole University of Technology, Opole, Poland

[2] Faculty of Engineering Production and Logistics, Opole University of Technology, Opole, Poland

[3] Faculty of Automatic Control and Informatics, Opole University of Technology, Opole, Poland