Application of data mining techniques to indoor and outdoor air studies

Dissertation

zur Erlangung des Grades

„Doktor der Naturwissenschaften“

im Promotionsfach Chemie

am Fachbereich Chemie, Pharmazie und Geowissenschaften

der Johannes Gutenberg-Universität Mainz

Christof Stönner

geb. in Frankenthal

Mainz, den 30.05.2018

Summary Humans emit a wide range of volatile organic compounds (VOCs). These molecules can be emitted via breath and skin and can be from endogenous or exogenous sources. The main breath gases besides N_2 and O_2 include CO_2, acetone and isoprene and are mainly endogenously produced via metabolic pathways. Exogenously emitted molecules comprise methanol from the digestion of fruits and molecules such as monoterpenes and siloxanes used in hygiene products. The study of these human-made emissions is important for the detection of biomarkers for illnesses as well as for the estimation of the contribution of human emission to indoor and outdoor environments. The measurement of volatile organic compounds in indoor and outdoor studies was performed with a proton transfer reaction time-of-flight mass spectrometer (PTR-TOF-MS).
Closed spaces with controlled ventilation such as the showroom of a cinema allows the estimation of emission rates from a large group of people averaging over individual behaviour and habits. Factors such at diet or use of hygiene products depict the largest source for uncertainty in estimating the emission rates. On a much smaller scale the emission of human-emitted molecules varies with the emotional state. In the cinema showroom the screening of a film induces the same stimuli on a large amount of people and reproducible patterns in the time series of VOCs were found. The combination of the measured time series of VOCs and film scene annotations and the application of data mining techniques allows the discovery of relationships between the emission of VOC and specific scenes displayed in the film.
Most of the world population now lives in urban areas and humans spend most of their time in indoor environments. In closed spaces people are exposed to volatile organic compounds which can occur in much higher abundances than outside. Since some of the VOCs can have adverse health impacts on humans it is important to estimate sources of VOCs in indoor environments such as emissions from furniture, human emissions and VOCs being transported from outside into these closed spaces.
These outside sources are strongly dependent on biogenic sources such as emission of plants and vegetation and anthropogenic sources for example through combustion processes. Human emission can significantly impact the air chemistry in urban areas but on a global scale they only contribute a small amount to the total emission of VOCs. The behaviour and fate of a VOC is affected by many factors such as temperature, relative humidity and the origin of the air mass. To study the atmospheric chemistry of these VOCs, measurement campaigns were conducted in different locations lasting over 4 weeks. Typically, different meteorological conditions are faced during this measurement period. In order understand the atmospheric behaviour of a VOC it is useful to partition these time series in periods of similar meteorological conditions. To do this objectively a pattern identification method was applied. The data-driven investigation of the time series provided useful insights in the chemistry behind the VOCs. The proton transfer reaction time-of-flight mass spectrometer is able to capture hundreds of VOCs in real time and therefore the combination of this instrument with data mining techniques has huge potential for future research projects.

Zusammenfassung Menschen emittieren eine Vielzahl von flüchtigen organischen Verbindungen (VOCs). Diese Moleküle können über den Atem und die Haut emittiert werden und stammen aus endogenen oder exogenen Quellen. Die wichtigsten Atemgase neben N_2 und O_2 sind CO_2, Aceton und Isopren und werden hauptsächlich endogen über Stoffwechselwege produziert. Exogen emittierte Moleküle schließen Methanol aus der Verdauung von Früchten und Molekülen wie Monoterpene und Siloxane aus Hygieneprodukten ein. Die Untersuchung dieser menschlichen Emissionen ist wichtig für die Identifikation von Biomarkern von Krankheiten sowie für die Abschätzung des Beitrags menschlicher Emissionen zur Innen- und Außenluft. Die Messungen von flüchtigen organischen Verbindungen in der Innen- und Außenluft wurde mit einem Protonentransfer-Flugzeitmassenspektrometer (PTR-TOF-MS) durchgeführt.

Durch die Messung von menschlichen Emission in geschlossene Räumen mit kontrollierter Ventilation zum Beispiel in einem Kinosaal können Emissionsraten von einer großen Gruppe von Menschen, die sich durch individuelle Verhaltensweisen und Gewohnheiten unterscheidet, berechnet werden. Die größte Unsicherheit bei der Schätzung dieser Emissionsraten stellen Faktoren wie Ernährung oder Konsum von Hygieneprodukten dar. In einem viel kleineren Maßstab variiert die Emission von menschlichen Molekülen mit dem emotionalen Zustand. Durch das Zeigen eines Films werden die gleichen Reize bei einer großen Anzahl von Menschen induziert und reproduzierbare Muster in der Zeitreihe der emittierten Moleküle wurden beobachtet. Die Auswertung der gemessenen Zeitreihen und der gezeigten Filmszenen mithilfe von Data Mining-Methoden zeigte Korrelationen zwischen menschlichen Emissionen von VOCs und Filmszenen.

Ein Großteil der Erdbevölkerung lebt in städtischen Gebieten und die Menschen verbringen die meiste Zeit in Innenräumen. Dort sind die Menschen VOCs ausgesetzt, welche gesundheitsschädlich sein können und zum Teil in höheren Konzentrationen vorkommen. Daher ist es wichtig, Emissionen von VOCs in Innenräumen wie zum Beispiel aus Möbeln sowie menschliche Emissionen und VOCs, die von außen in diese geschlossenen Räume transportiert werden, zu charakterisieren.

VOCs in der Außenluft sind stark abhängig von biogenen Quellen wie der Emission von Pflanzen und Vegetation und anthropogenen Quellen durch Verbrennungsprozesse. Die Emissionen von Menschen können die Luftchemie in städtischen Gebieten erheblich beeinflussen, aber auf globaler Ebene tragen sie nur einen kleinen Teil zur Gesamtemission von VOCs bei. Das Verhalten eines Moleküls wird von vielen Faktoren wie zum Beispiel Temperatur, relative Luftfeuchtigkeit und der Herkunft der Luftmasse beeinflusst. Um die Zusammensetzung dieser VOCs in der Atmosphäre zu untersuchen, wurden an verschiedenen Standorten Messkampagnen über 4 Wochen durchgeführt. In der Regel sind während dieser Messperiode unterschiedliche meteorologische Bedingungen zu beobachten. Um das Verhalten von VOCs zu verstehen, ist es sinnvoll, diese Zeitreihen in Zeiträume ähnlicher meteorologischer Bedingungen zu unterteilen. Dafür wurde eine datengesteuerte Musteridentifikationsmethode angewendet. Die Untersuchung der Zeitreihen lieferte nützliche Einsichten in die Chemie hinter den VOCs. Das PTR-TOF-MS ist in der Lage hunderte Moleküle in Echtzeit zu messen und daher bietet die Kombination von Data Mining-Methoden ein großes Potential für die Auswertung dieser Daten.

Contents

1. Introduction

In this chapter the basic information required for the understanding of the proton transfer reaction time-of-flight mass spectrometer (PTR-TOF-MS) and the applied data mining techniques are presented. First the set-up and operating principle of the PTR-TOF-MS are described. The second section introduces the applied data mining techniques and the methods for model evaluation.

1.1. Proton transfer reaction time-of-flight mass spectrometer (PTR-TOF-MS)

The proton transfer reaction time-of-flight mass spectrometer is widely used as an instrument in the field of atmospheric chemistry and for the investigation of human volatiles. This analytical tool allows the online measurement of numerous volatile organic compounds (VOCs).[12, 57, 66] The final data generated by this method is typically reported at a time resolution of about one minute (down to 0.1 seconds) and contains hundreds of masses up to a mass-to-charge limit of approximately 400 m/z. The mass resolution ($m/\Delta m$) is in the range of 4000-5000 allowing the separation of isobaric masses. The reasons leading to this characteristic data is discussed in the following sections.

1.1.1. Setup of the PTR-TOF-MS

The setup of the PTR-TOF-MS is shown schematically in Figure 1.1.

On the left side in Figure 1.1 the hollow cathode generates hydronium ions (H_3O^+) in a glow discharge. These reagent ions are transported using an electromagnetic field into the drift tube through which the air containing the analytes is drawn. The analytes undergo a proton transfer reaction with the hydronium yielding positively charged analyte ions of the mass M+1 (original nominal mass plus a proton). The analyte ions are transferred out of the drift tube into a lens system that focuses the ion beam. In the time-of-flight section the ions within the continuous ion beam are accelerated orthogonal to the direction of the beam. Once the ions are extracted into the time-of-flight area the ions are separated according to their mass-to-charge ratio. This happens because every ion is initially accelerated with the same energy. The energy is transferred to each ion with an electric pulse, pushing the ions into the time-of-flight tube towards the detector. The transferred energy becomes kinetic energy dependent on the mass of the ion m and its velocity v ($E = \frac{1}{2}mv^2$). Since different molecules possess different molecular masses the velocity of an ion will vary too. Higher masses tend to be slower

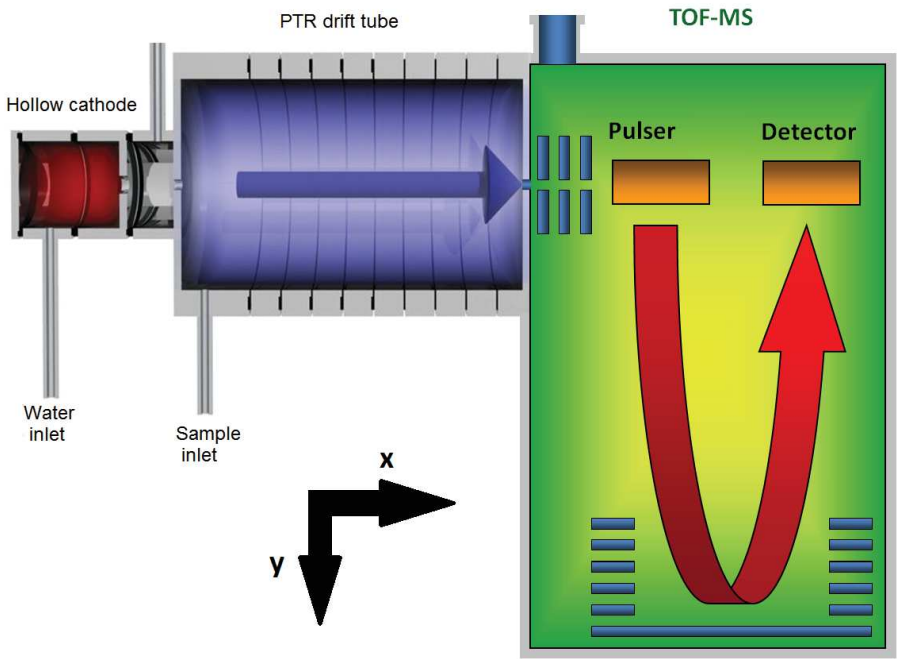


Figure 1.1.: Set-up of the PTR-TOF-MS.

in the flight tube. Therefore, the molecules reach the detector (multi-channel-plate) at different times which are precisely recorded. When impacting the detector, electrons are released which are amplified and finally a current can be measured.
The extraction of the ions from the continuous ion beam into the time-of-flight area takes place every 0.1 nanoseconds. The detector counts the impacting ions and calculates the required time from the extraction to the impact. Depending on the desired time resolution (typically ranging from 0.1 seconds to several minutes) the counting statistics for each extraction is summed up resulting in a single mass spectrum. The frequency of the extraction defines the upper mass-to-charge limit. The higher the frequency the lower the mass-to-charge limit since at some point the heavier compounds (slower velocities) are not able to reach the detector in time before the next extraction.

1.1.2. Proton affinity

The air with the analytes is drawn in the drift tube where they undergo a proton transfer reaction with the hydronium ions. This reaction only takes place if the proton affinity (PA) of the analyte is higher than for water (H_3O^+). Table 1.1 shows a few selected compounds and their corresponding proton affinity.
Water has a proton affinity of 691 kJ mol^{-1}. The main constituents of air such as nitrogen, oxygen, argon and carbon dioxide have a lower proton affinity than water and are not protonated in the drift tube. Also, many hydrocarbons like alkanes possess a lower proton affinity than water and are undetectable. Components being protonated include

many alkenes like isoprene, aromatics, oxidated species like alcohols, aldehydes and organic acids. Furthermore compounds containing sulfur and nitrogen can be measured such as dimethylsulfide and acetonitrile. This chemical ionization is considered as a soft ionization method as the relatively low excess energy transferred to the analytes results in low fragmentation of the molecules.

Table 1.1.: Proton affinities of selected compounds.

Group	Compound	Proton affinity [kJ mol^{-1}]
Inorganic gases	O_2	421
	N_2	494
	CO_2	541
	H_2O	**691**
	$(H_2O)_2$	808
Alkanes	Methane	544
	Ethane	596
Alkenes	Ethene	641
	Propene	752
Alkines	Acetylene	641
Aromatics	Benzene	750
	Toluene	784
	Phenol	817
organic compounds	Chloromethane	647
	Formaldehyde	713
	Acetone	812
	Ethanol	776

1.2. Data mining

In general, the data mining task can be divided into supervised and unsupervised methods. On the one hand, supervised learning comprises tasks where the output class labels are already known to the learner. The known variable is also called "dependent variables" whereas the variables used to predict this variable are called "independent variables". The supervised methods are used to predict the dependent variable by identifying structure in the independent variables. The methods can be divided up into models which can be easily interpreted like linear and logistic regression and decision tree models or into "black-box models" which are difficult to interpret such as random forests, neural nets and support vector machines. Which of these methods should be applied depends on the data and on the aim of the analysis. Simpler models are not able to capture non-linear behaviour in the set of independent variables. If such a structure exists in the data a more complex model must be used. On the other hand, unsupervised learning deals with the aim of dividing up the data into a number of groups. In this case, the

class labels of the data are not known to the user. Unsupervised learning methods comprise clustering methods like hierarchical clustering and dimension reduction methods. Additionally, a sequence mining algorithm is presented for finding patterns in discrete temporal data.

1.2.1. Supervised Methods

This section provides the basics of model validation and presents two classifiers namely decision tree and random forest models. An example data set can be seen in Table 1.2 showing some data from a measurement campaign in Cyprus. The data was put into two different classes labelled as "high" and "low" being the dependent variable. The rest of the data are independent variables (temperature, relative humidity, wind speed, wind direction). Each row is called an instance (x, y) defined by a dependent variable x and a set of independent variables y. For a classification task the dependent variable must be discrete otherwise it is referred as regression. The supervised method tries to map the independent variables with the help of a target function f on the dependent variable.
If the user is only interested in explanatory data analysis the whole data set is used and

Table 1.2.: Example data set including a binary dependent variable and 4 independent variables (temperature, relative humidity, wind speed and wind direction).

Dep. variable	Temperature	Rel. humidity	Wind speed	Wind direction
high	24.8	87.3	1.35	SW
high	25.0	85.6	0.9	SSW
high	25.1	72.5	2.75	SW
high	25.2	78.4	2.5	SW
low	25.5	73.7	3.15	SSW
low	25.7	68.8	2.3	SSW
high	25.9	65.9	2.5	SW
low	26.1	69.5	1.6	SW
low	26.4	65.4	2.5	SW
low	26.5	59.1	2.45	SSW
low	27.0	59.0	3.15	SSW

the representation of the classification model is interpreted. For example, the slope of a linear regression can be used for further interpretation. In predictive analysis the whole data set is divided into a training set usually containing two thirds of the data and into a test set containing one third of the data. The training data is used to build the model and the test set is used to predict the target class from each unseen instance. The whole procedure is depicted in Figure 1.2. The outcome of the prediction is compared to the true classes of the test set to evaluate and validate the model.[54, 98, 168]

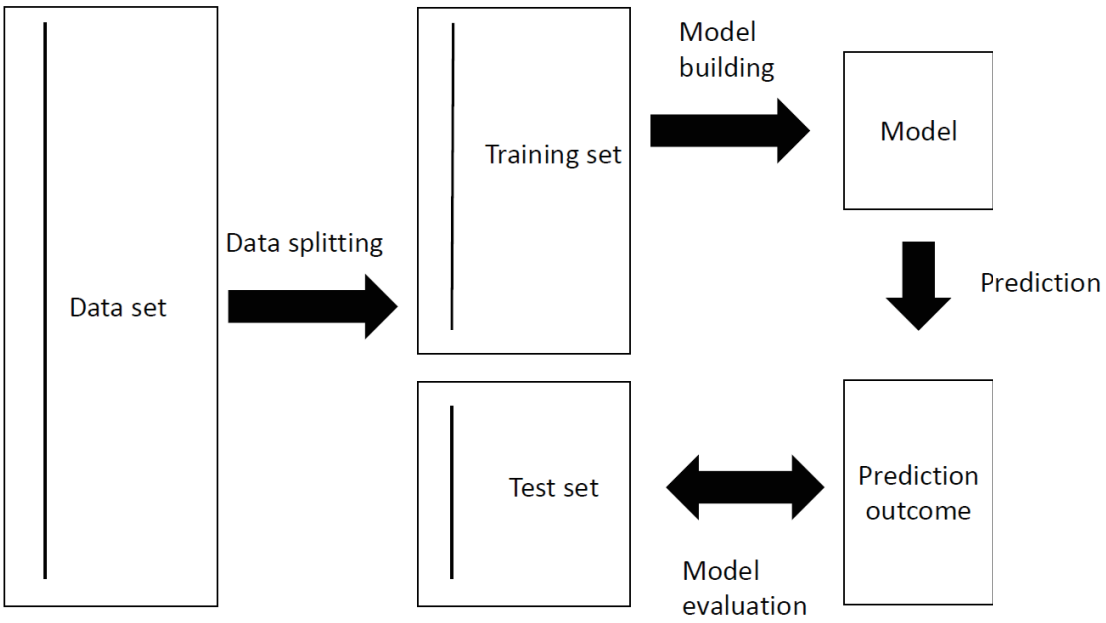


Figure 1.2.: Scheme for model building and evaluation.

Model validation

This section discusses methods for model evaluation being an essential part in the development of a model. Model evaluation helps to find the model which represents the given data best. Therefore, some performance measures must be established being able to evaluate the prediction outcomes of the model. Using the whole data as a training set in order to build the model would lead to overoptimistic and overfitted models. The meaning of overfitting in conjunction with techniques tackling this problem are also presented in this section.

For the model evaluation several measures can be taken. Figure 1.3 shows the layout of a confusion matrix. Each column represents the instances of the actual class and each row the instances with their predicted class. It is usually presented in conditions labelled as "TRUE" and "FALSE" for a binary variable. The four fields are labelled as "true positives" for instances predicted as the actual class, "false positives" for instances that were predicted as "TRUE" with an actual class of "FALSE", "false positives" for instances predicted as "FALSE" and an actual class of "TRUE" and "true negative" for instances that were correctly predicted as "FALSE". These four terms are used to derive performance measures such as *Accuracy*.

$$Accuracy = \frac{\sum \mathrm{TP} + \sum \mathrm{TN}}{\text{total population}} \tag{1.1}$$

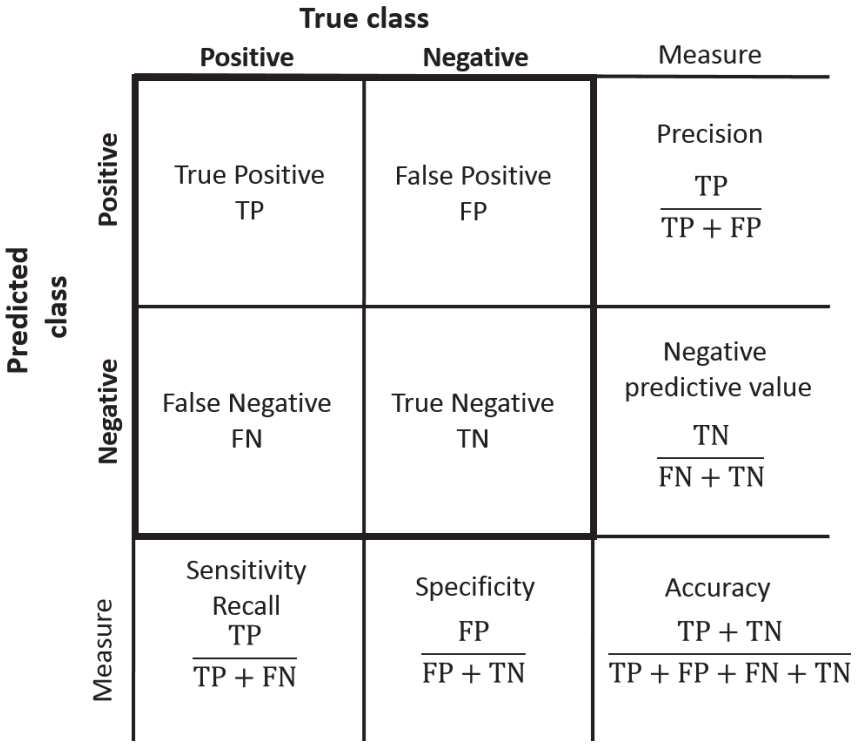


Figure 1.3.: Confusion matrix with several performance measures.

This measure adds up the correctly predicted instances divided by the total number of instances. However, this measure becomes unreliable if the data gets unbalanced with classes that differ greatly in number. For example, if there are 95 sample of the label "TRUE" and 5 of the label "FALSE" and all of them were predicted as "TRUE" the model would result in 95% accuracy even if the model classifies all samples just as "TRUE". Therefore, other performance must be used if the data is highly unbalanced. A useful performance measure can be Cohens's kappa κ when it comes to unbalanced data sets.

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{1.2}$$

in Equation 1.2 p_0 is the relative observed agreement (same as *Accuracy* in Equation 1.1) and p_e is the hypothetical probability of chance agreement and is defined as following:

$$p_e = \frac{\text{TP} + \text{FP}}{\text{total population}} \cdot \frac{\text{TP} + \text{FN}}{\text{total population}} + \frac{\text{TN} + \text{FN}}{\text{total population}} \cdot \frac{\text{TN} + \text{FP}}{\text{total population}} \tag{1.3}$$

Cohen's kappa lies always between 0 and 1 and basically tells how much better the model performs than random classification.[21]
Until now the model predicts the target label of an instance. For some classifiers such as randomForest models it is possible to derive the probability of how likely it is that one instance belongs to a certain label. The sum of the probabilities must sum up to 100%.

These probabilities can be used to define a threshold value classifying the instances into positively (TRUE) and negatively (FALSE) predicted outcomes. This threshold value is usually set to 0.5, yielding a positive outcome if the probability is larger than 0.5 and otherwise resulting in a negative outcome. However, it is not always obvious how the right threshold value should be chosen. Therefore, performance measures such as receiver operating characteristics (ROC) curves and precision-recall-curves (PRC) are introduced providing an overview over the whole range of possible threshold values. The ROC plot shows the trade-off between the false positive rate, which is 1- specificity (1 - SP = TN / (TN+FP), and sensitivity (SN = TP / (TP+FN) as shown in Figure 1.4. In case of the ROC curve the diagonal line from the origin (0,0) to the point (1,1) would be a random classifier and a curve above of this straight line depicts a better prediction than just by chance. A single performance measure can be obtained by calculating the area under the ROC curve (AUC). For a random classifier this would result in a value of 0.5 and for a perfect classifier in a value of 1.[37, 38] For unbalanced data sets the PRC is a better choice.[118] This plot shows the trade-off between precision (PREC = TP / (TP +FP)) and recall (REC = TP / (TP + FN).

In predictive analysis cross-validation is used to reduce the problem of overfitting. This

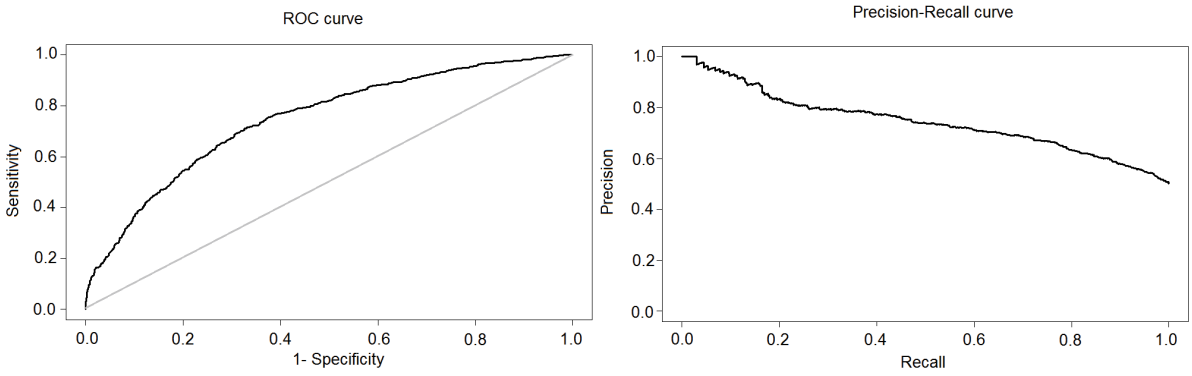


Figure 1.4.: Representation of a ROC curve (left side) and a precision recall curve (right side).

happens if the model tries to find patterns in the training data and may pick up some patterns that are just caused by random chance rather than by true properties of the unknown function f. For example, for a linear model it is possible to choose a function of the polynomial degree of one or to include higher order polynomials. If the order of the polynomial is equal to the number of instances, the function can perfectly capture the behaviour of the instances. However, this describes the problem of overfitting since the function f perfectly fits to the training data but performs poorly on unseen test set data. Thus re-sampling techniques such as cross-validation are used to obtain the optimal set of parameters given to the model. Taking the example of a linear model, a suitable polynomial degree can be found. The idea is to divide the training set, which is used for model building, internally into additional training and test sets. These sets of the original training set are used to choose the model which performed best with a given set of parameters.[81] These parameters are used to build the final model using

the whole training set. The original test set should not be incorporated in the model building process and is used as an unseen set of instances for the final model.
A procedure to evaluate whether the model really captures some pattern in the training set data or finds some random structure is the use of permutation tests. First a model is trained on the normal data set and a performance measure is calculated. Then the classes of the dependent variable are randomly permutated and again a model is trained and evaluated. This permutation is repeated several times. Then the performance measures of the permutated model are compared to the original data and the occurrences how often the performance measure of the random model is greater than the performance measure of the original model is recorded. This fraction indicates if the original model found some real pattern in the data.[108]

Decision trees and random forest models

This section provides the basics for the understanding of decision trees. Figure 1.5 shows a possible representation of a decision tree of the fictional data included in Table 1.2.

The decision tree represents a hierarchical structure starting at the root node and

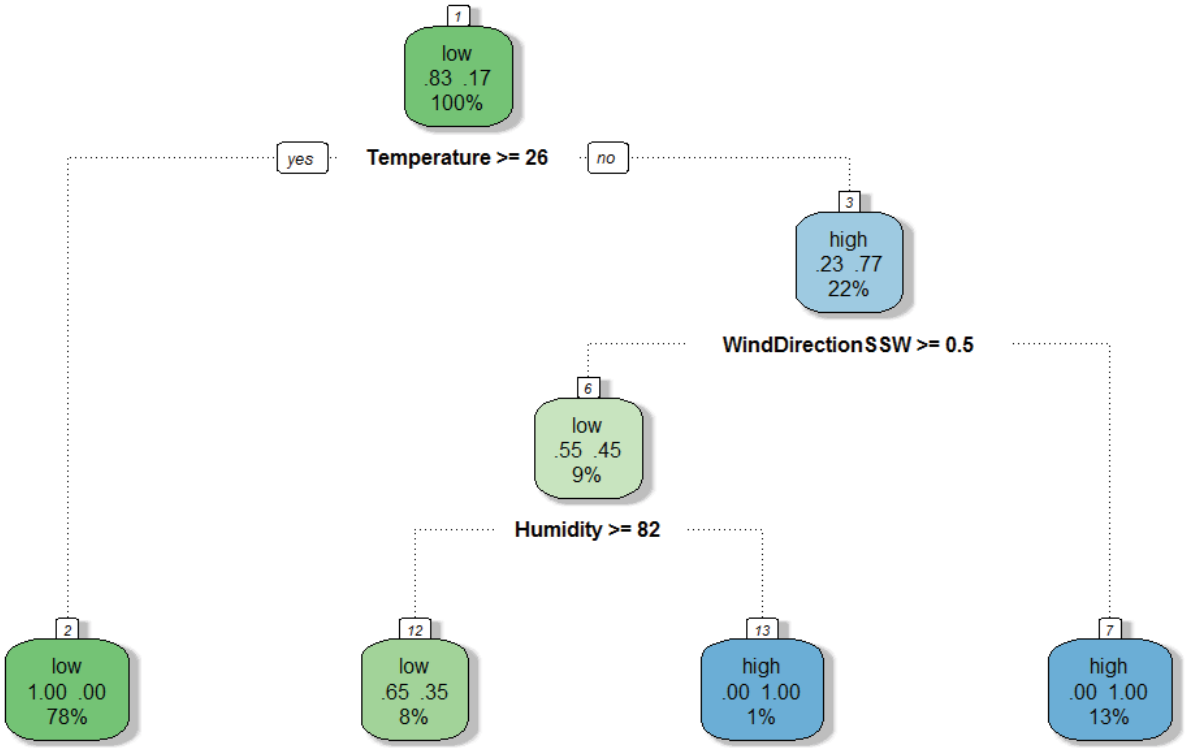


Figure 1.5.: Possible representation of a classification tree model for the example data in 1.2.

leading over internal nodes to several leaf nodes. At each non-terminal node (including the root and internal nodes) a conditional test is performed on one of the independent variables. Figure 1.5 shows a representation of a decision tree model. The top node is

called root node and the bottom nodes are called leaf nodes. In the boxes, the lowest line shows the amount of data included in that node from the training set. The root node contains all the data (100%) and the leaf nodes contain some fraction of this data. For example box number 2 (left bottom box) contains the data for which the temperature is greater or equal than 26 which is the case for 78% of the data. The second line in the box shows the distribution of the classes for the data contained in this node. For the root node containing all the data the initial distribution of the class labels is 83% of the label "low" and 17% of the label "high". The majority is shown in the first line stating that the label "low" is the most frequent label. In general, the notation of the splitting nodes is as following: "<variable name> <condition>". First the variable name is presented and the condition shows the splitting criteria. In general, the higher the nodes the more important their influence on the classification result. In case of explanatory analysis one can examine each node reflecting the variable importance and investigating the splitting criteria. For prediction a new instance is introduced at the root node and at each node the corresponding variable is tested whether the condition is TRUE or FALSE. If the condition is TRUE, the left path is taken otherwise the right one. This is done until this instance ends at some leaf node and the majority class of the model tree for this leaf node is given to the new instance.
Now the question arises in which order the independent variables are selected and how the test condition is calculated. Therefore, a statistical property measure of how well a given independent variable A separates the data (S) into the target classification (given by the dependent variable) must be established. This property is called *information gain*. In order to calculate the *information gain* of a dependent variable, we first define the term *entropy* describing the disorder of a given distribution of discrete values.[123]

$$Entropy(S) = -p_{\text{high}} \cdot \log_2(p_{\text{high}}) - p_{\text{low}} \cdot \log_2(p_{\text{low}}) \tag{1.4}$$

Here we assume a binary variable (containing values of "high" and "low" from the dependent variable in Table 1.2). The variable p_{high} in Equation 1.4 describes the proportion of the class "high" and p_{low} the proportion of the class "low". For example, the data set consists of 11 instances with 6 of them belonging to class "high" and 5 of them to class "low". Inserting the proportions ($p_{\text{high}} = \frac{6}{11}$ and $p_{\text{low}} = \frac{5}{11}$) into Equation 1.4:

$$\begin{aligned} Entropy[\text{6high,5low}] &= -\frac{6}{11} \cdot \log_2(\frac{6}{11}) - \frac{5}{11} \cdot \log_2(\frac{5}{11}) \\ &= 0.994 \end{aligned} \tag{1.5}$$

For a binary variable with instances belonging to the same class, the *entropy* term is 0 whereas for a distribution of equal occurrences of the two classes it results in an *entropy* value of 1. For an unequal distribution of a binary variable the *entropy* values lies between 0 and 1. The function of *entropy* for a binary variable is shown in Figure 1.6.
The *entropy* term defined in Equation 1.4 leads to a measure which allows to estimate how well an independent variable separates the dependent variable in its different classes. This measure is called *information gain*. It describes the reduction in entropy of the

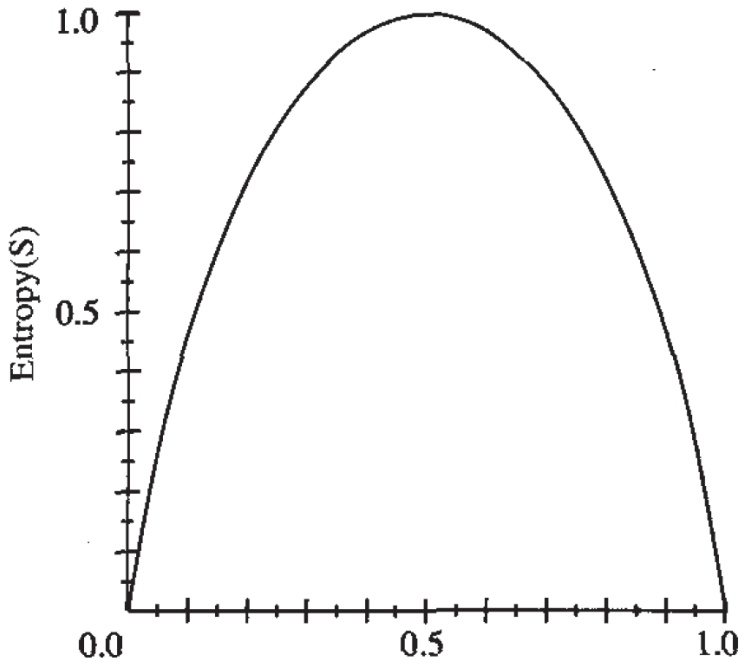


Figure 1.6.: The entropy function for a binary variable.[98]

dependent variable by classifying the data according an independent variable. This *information gain* or $Gain(S, A)$ is defined as following given the data (S) and the independent variable A:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \tag{1.6}$$

The term $Values(A)$ incorporates all possible values of the independent variable A and S_v is the subset of the data S for which the variable A has the value of v. The first term in Equation 1.6 is the *entropy* of the complete data set. The second term includes the *entropy* of the subset S_v being partitioned by value v of variable A. This *entropy* of the subset S_v is weighted by the size of this subset $|S_v|$ divided by the size of the whole data set $|S|$. Thus the subtraction of the original *entropy* by the sum of the *entropy* calculated from the partitioned data set results in the reduction in *entropy* by variable A. The *information gain* is used to evaluate the relevance of each variable and to select the best for growing the tree. For example, taking the data from Table 1.2 and calculating the *information gain* by choosing the wind direction variable for classification. The wind direction is a binary variable with values of "SSW" occurring $|S_{\text{SSW}}| = 5$ times and "SW" occurring $|S_{\text{SW}}| = 6$ times. In case of wind direction = "SSW" the target class is labelled "high" for one time and "low" for 4 times. On the other hand, if the wind direction = "SW" the target attribute shows 4 times the label

"high" and 2 times the label "low". Thus the Equation 1.6 is as following:

$$\begin{aligned} Gain(S,A) &= Entropy(S) - \sum_{v\in\{\text{SSW,SW}\}} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \\ &= Entropy(S) - \frac{5}{11} \cdot Entropy(S_{\text{SSW}}) - \frac{6}{11} \cdot Entropy(S_{\text{SW}}) \\ &= Entropy(S) - \frac{5}{11} \cdot Entropy([\text{1high,4low}]) - \frac{6}{11} \cdot Entropy([\text{4high,2low}]) \\ &= 0.994 - 0.455 \cdot 0.722 - 0.545 \cdot 0.918 \\ &= 0.165 \end{aligned} \tag{1.7}$$

This outline also works for dependent variables with more than two values but only for binary independent variables. If more than two values are present for an independent variable new dummy variables for each of these values must be created. This dummy variable contains the information about one selected value in this variable if it is present (the value for this instance is set to TRUE) or not (the value for this instance is set to FALSE). These new independent and binary variables are treated as before.
Taking the example from Figure 1.5 the *information gain* is calculated at each non-terminal node and the variable is selected for which the largest value is calculated. In the presented outline only categorical variables were allowed. For continuous variables the variable must be discretized beforehand. The continuous-valued variable is discretized into two classes based on passing a conditional test in the form of $A < c$. For smaller values of the variable A than the threshold value c the new discretized variable is set TRUE otherwise the variable is set FALSE. The threshold c is set by selecting the greatest *information gain*. This can be done by sorting the data according to the variable A and picking out adjacent instances which change their target classification. For all of these occurrences a candidate threshold value is chosen which lies in the middle between the two surrounding values. Then the *information gain* for all chosen threshold candidates is computed and the greatest value in *information gain* is selected as a threshold for this variable. Finally, the calculated *information gain* is compared to the rest of the independent variables.[98, 113] For example, Table 1.2 was sorted according to an increase in temperature. Three possible splits are available at temperature values of 25.6, 25.8, and 26.0. Transforming the temperature variable into a binary variable by dividing the temperature in values greater and smaller than the splitting criteria results in *information gain* values of 0.311, 0.165 and 0.445. Thus the third splitting criteria (Temperature > 26) is chosen. Compared to the *information gain* value of the wind direction variable of 0.165 the temperature variable with a splitting criteria of "Temperature > 26" is chosen to be the first splitting node (the two other variables show lower *information gain* values, too).
This presented procedure for applies for classification trees when the dependent variable is discrete. For regression trees with a continuous dependent variable first the independent variables are divided into non-overlapping regions. This is done by recursive binary

splits. This procedure is the same as for the classification tree outlined in the above paragraphs. For each possible independent variable and each possible splitting value (discrete or continuous) the residual sum of squares (RSS) of the dependent variable must be minimized (similar to maximizing the $information\ gain$ for classification). A discrete independent variable A divides the data into two subsets belonging to one of the values (for example $S_1(A, \text{high})$ and $S_2(A, \text{low})$ in Equation 1.4). In case of a continuous independent variable A the cut point c dividing the dependent variable in partitions of $A > c$ and $A <= c$ is selected which minimizes the RSS. The two resulting subsets (dividing the independent variable into subsets one larger and the other smaller or equal than c) are labelled as $S_1(A, c)$ and $S_2(A, c)$. For each of these subsets the corresponding mean value of the dependent variable $\hat{y}_{S_1}$ and $\hat{y}_{S_2}$ is calculated. The RSS is defined as following:

$$RSS = \sum_{i \in S_1(A,c)} (y_i - \hat{y}_{S_1})^2 + \sum_{i \in S_2(A,c)} (y_i - \hat{y}_{S_2})^2 \tag{1.8}$$

Thus for each subset the sum of the difference between each value of the dependent variable y_i belonging to this subset and its mean value is calculated. This value is added to the second subset. This is done for each independent variable and each possible cut point or in case of a discrete variable its values. The variable and cut point is chosen as a splitting node which minimizes the RSS. The subsequent splits are performed as for the classification tree on the subsets S_1 and S_2. For an unknown instance the mean value of the subset (one of the leaf nodes), to which this unseen observation belongs, is returned.
To reduce the overfitting for these models some pruning of the trees is done. Therefore, a cut-off value must be defined balancing the tree size with the goodness of fit.[80, 114] The cut-off value is usually chosen via internal cross-validation. Figure 1.7 shows an example comparing the accuracy of the training set with the one of the test set by increasing the number of nodes. The optimal value for the number of nodes can be found by cross-validation. It can be seen that the accuracy of the training set increases monotonically whereas the accuracy of the test set first increases and then decreases by adding further nodes. This is because by adding further nodes no real pattern in the data is found and the splits are made only due to the noise in the training set.

The next algorithm used in this thesis is a randomForest classifier and it is based on the decision tree algorithm. In the case of randomForest models multiple trees (typically around 1000 trees) are grown. These trees are built upon a sub-sample of the whole data set such that some instances are omitted and the number of variables are restricted for each tree. For example, for the first tree 33% of the instances were randomly removed and out of 10 variables only 5 variables were allowed for model building. For the second tree other instances and variables were randomly removed. The final prediction combines all results from the trees and the predicted label is the majority vote from the output of each tree.[16] This procedure results in probabilities for each instance belonging to one class (for example 70% out of the 1000 trees predict this instance belonging to class “A” and the rest to class “B”). These probabilities can be used for ROC curves and PRC. Compared to decision tree models this leads to a larger stability of the model.

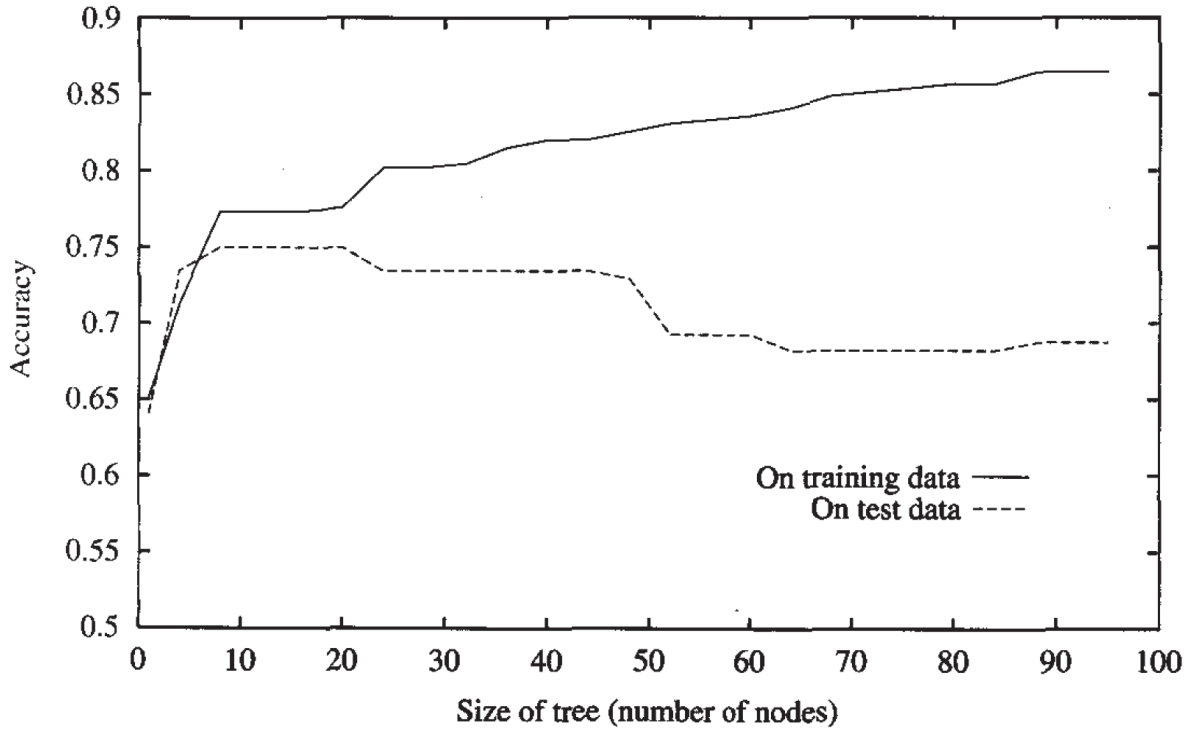


Figure 1.7.: This shows the effect of overfitting for a decision tree. The number of nodes is plotted against the accuracy of the training set and test set. It can be seen that the accuracy for the training set steadily increases with increasing number of nodes whereas the accuracy of the test set first increases and then decreases reaching its maximum around 10 nodes.[98]

In the case of the decision tree a small change in the values in one of the variables can alter the way the first split is done and subsequently the representation of the whole tree can be altered. Thus the representation is highly dependent on the given data. To tackle this problem, the random forest model introduces some diversity. This is done by subsampling the data (removing some of the instances) and by omitting some of the variables. This procedure reduced the chance of overfitting through the internal re-sampling methods. This becomes very important if there are a lot of independent variables or if the user wants to extrapolate the model to new data. Random forest classifiers are seen as robust classifier compared to other classifiers because of their internal re-sampling method and variable selection process. Due to their implicit variable selection they can also handle data set with many variables obtaining no signal but only noise. These kind of variables do not add any further information to the classification problem. Therefore, randomForest classifier are suitable for extrapolation to a different set of instances.

1.2.2. Unsupervised Learning

In unsupervised learning the goal is to partition the data set into the number of desired clusters or to mine frequent patterns in discrete data. This task can be performed by hierarchical clustering and sequence mining. These two algorithms are presented in the following sections.

Hierarchical cluster

A representation of a hierarchical clustering is shown in Figure 1.8. This representation is also called dendrogram. At the bottom of the dendrogram each variable is in its own cluster whereas the top node joins all variables into one cluster. At intermediate levels one may find meaningful groups.

To create such a dendrogram from atmospheric time series one needs to calculate

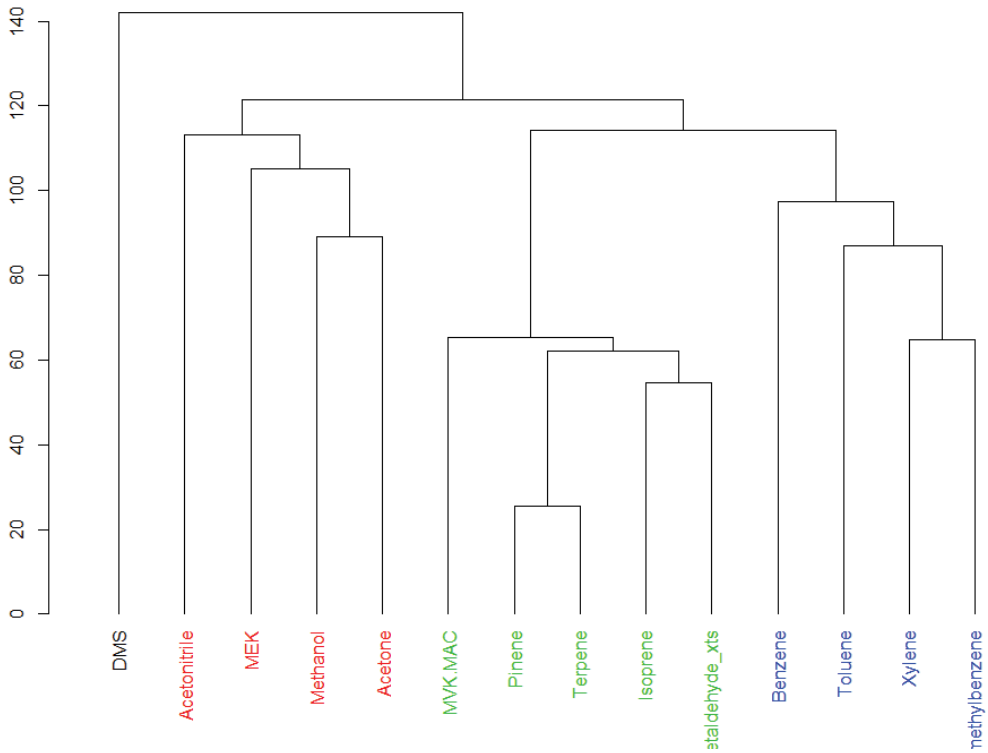


Figure 1.8.: Representation of hierarchical clustering result as a dendrogram.

the pairwise distance for each variable. The distance metric can be chosen by the user including for example Euclidean distance or correlation. This distance matrix serves as the input for the hierarchical clustering algorithm. Here we use agglomerative clustering. This method starts from the bottom with each variable in its own cluster. Then it merges successively the most similar pair of clusters until all clusters are in the same cluster

reaching the top node. At the very beginning the algorithm merges the two cluster (at the very beginning these are two variables) with the smallest distance. Doing this a new cluster (C_{ij} is formed containing two clusters C_i and C_j. For the next step the closest pair of clusters is found subsequently a new cluster is formed until there is only one cluster. When more and more variables are joined into one cluster the mean distance between two cluster must be calculated in order to find the closest connection.[71]

Sequence mining

Sequence mining is the task of discovering patterns in a discrete and temporal data set. Therefore, the time series must be divided into several segments for example days. The SPADE algorithm records the position of each label in each segment. Next it is checked for each day whether there is a label which occurs temporally after the chosen label. This is performed for each label and the found sequences of length two are recorded. These records of length two include the position of the second label which occurs after the first one and subsequently the next label can be added. For example, given the label "low" occurring at day 1 at positions 2,4 and 5 the following tuple will be created $< 1, \{2, 4, 5\} >$. Figure 1.9 shows an example of the labels "low" and "high" and their positions.

In order to merge these two labels creating the sequence of the labels from "low" to

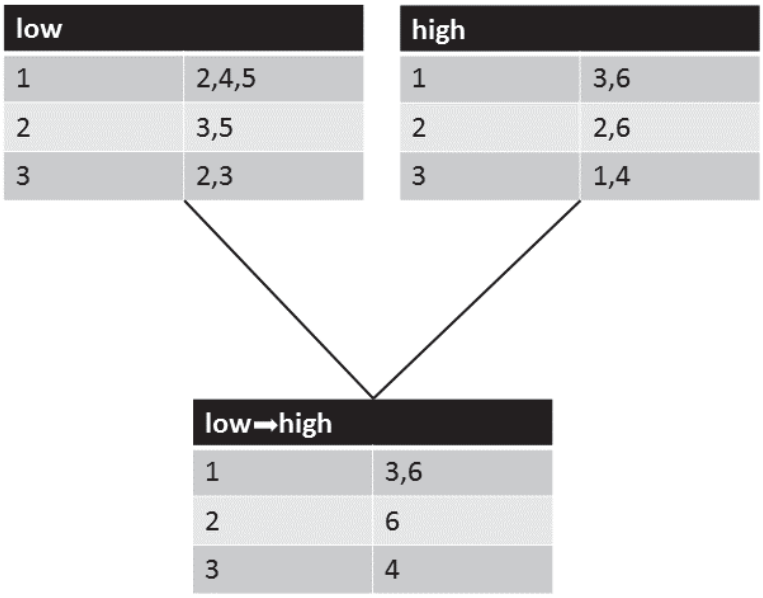


Figure 1.9.: Example of the SPADE sequence mining algorithm.

"high" for each day the position of the second label is tested whether it occurs temporally after the first one (larger position number). For the first day both the positions 3 and 6 for the label "high" occur after some occurrence for label "low", for example after

position 2. Thus for the first day the position {3,6} is recorded. Then further labels are added. The number of rows of the boxes in Figure 1.9 represent the total occurrences of this label or sequences in the data set. The fraction of these rows to the total number of segments (here days) is called *support*. If the *support* for one sequence is smaller than a predefined threshold value this sequence will be dropped and no further labels will be added. Furthermore, the maximum distance between two labels can be defined. For example, if the addition of a new label must occur subsequently to the former one with the distance of one or if larger gaps are allowed.[167, 168]

1.3. Open research questions

The PTR-TOF-MS finds its ideal applications in domains where rapid changes of the measured air is expected. Here, two application fields are presented comprising the measurement of human emissions and the measurement of VOCs in the atmosphere.

1.3.1. Human emissions

The PTR-TOF-MS is widely used in the field of measuring human emissions. One application area involves the measurement of human breath in order to identify biomarkers and to establish a non-invasive method for detecting illnesses like lung cancer.[6, 111, 115] Furthermore, the influence of groups of people on the indoor air chemistry has been investigated. The settings for these measurements took pace in office rooms,[126] class rooms,[117, 135, 136, 144] football stadiums[142] and as in our case the showroom of a cinema. These measurements bear the advantage of estimating emissions from large groups and thus averaging over individual habits and backgrounds. The PTR-TOF-MS is well suited to measure these human emissions since they can change rapidly due to opening or closing doors, dress and undressing clothes or also because of physiological changes like the increase of the breathing rate. Until now more than 1800 organic compounds are known to be emitted from humans.[24] The following bullet points summarize the open research questions.

- Veres et al.[142] investigated the emission of VOC from humans during a soccer match in a football stadium and the emission of these VOCs were extrapolated to global scale. It was discovered that ethanol was the dominant VOC besides the emission of CO_2. Additionally, tracers of smoking and skin ozonolysis were found. It can be seen that the emission of VOCs depends on many complex confounding factors such as age, dietary habits, consumption of hygiene products.[41, 77, 91] The measurement of larger groups and during several times of day is required to average over individual behaviour and to assess confounding factors influencing the human emission of VOCs. Therefore, the measurement in an enclosed environment with many people such as in a cinema showroom allows to study of these factors. The confined space and the steady flush rate of air allows the quantification of

the emission rates of VOC from humans and to examine the difference between children and adults as well as between endogenously and exogenously produced VOCs.

- The study performed by Veres et al.[142] in the football stadium also shows that through ozonolysis of skin lipids new VOCs were formed. A prominent candidate for these processed compound is 6-methyl-5-hepten-2-one (6-MHO).[149, 156, 157] The impact of the environmental condition such as ozone mixing ratio, temperature and abundance of rainfall can be attributed to the emission of human-borne VOCs. For indoor air studies these molecules can be transported into closed spaces and released into the air.

- During the measurement in the football stadium no goal was scored and the match resulted in a draw between the two teams. Nevertheless, the question arises if it is possible to see a goal in the measured time series of the VOCs specifically if the emission rates of human-borne VOCs are influenced by a goal

- The scoring of a goal certainly induces many emotions in the spectators. Enhanced emissions of VOCs might come from increased movement or from other physiological responses to this specific event. The influence of the emotional state might contribute to varying VOC levels in breath which could be an important factor in the search of biomarkers. The cinema provides a suitable environment to study emotions because the same emotional stimuli are exerted to all viewers. Additionally, many different emotions are induced through the showing of different scenes such as “action”, “comedy”, “drama”.

- A film consists of many different scenes inducing different emotions. It is unknown if the physiological response to these scene during a screening of a film can be used to classify a film into its age rating.

1.3.2. Atmospheric chemistry

The emission of the human-borne VOCs do not play a significant role in global atmospheric chemistry but can influence the local air chemistry.[142] A much larger source of VOCs is vegetation.[73] Although smaller in terms of carbon emitted, anthropogenic VOCs have an huge impact on the air chemistry.[25, 102, 170] These primary emitted compounds undergo various chemical transformations forming new species. The most important one is the oxidation through the hydroxyl radical (OH) and ozone (O_3) during daytime and the nitrate radical (NO_3) during night. The primary or secondary VOCs can form or adsorb to particles and surfaces (dry deposition) or can be washed out through rain (wet deposition).[9, 122] In some cases the VOCs may generate new particles.[8, 49, 61]
Many applications of data mining within the field of atmospheric chemistry involve the forecast of pollutants such as ozone, nitrous oxide and nitrous dioxide and particulate matter in urban areas. The data often consist of continuous measurements over years

for the main meteorological parameters and the main pollutants.[13, 20, 94, 106, 109] The data we use comprise many compounds (ca. 100 compounds) measured by various devices for a time period of around one month.

- The abundance of VOCs is influenced by various meteorological conditions such temperature, relative humidity, wind speed, wind direction and the origin of the air mass. Due to different conditions the behaviour of the VOC mixing ratio can change. There has not been an objective method for the segmentation of meteorological parameters of similar behaviour. Therefore, an objective segmentation of the measurement time period allows the comparison of the behaviour of the VOC under different conditions. Potential changes in the meteorological conditions can be the onset of the sea breeze or the decrease of the boundary layer height and thus the immersion into the free troposphere. Furthermore, the incorporation of VOCs with these meteorological variables can add valuable information for the segmentation of time series.

- Many methods are known for grouping VOC to specific groups such as principal component analysis, dimension reduction methods[161, 166] and clustering methods[63]. A new method is developed to cluster time series of VOCs into groups and to capture subtle differences between VOCs.

- Due to changing meteorological conditions it is difficult to estimate the influence of one variable on the behaviour of a VOCs. The interpretation of data mining methods can help to gain insights into the impact of a meteorological variable and help to understand the origin and fate of a VOC.

This thesis is divided up into an indoor air chemistry part and into an atmospheric chemistry part. First the estimation and interpretation of human emission rates are presented (chapter 2-3). Then the relationship between emotions and human emission is investigated (chapter 4-6). The second part focuses on the identification of patterns in atmospheric time series and the application of data mining methods for further understanding (chapter 7).

2. Real world volatile organic compound emission rates from seated adults and children for use in indoor air studies

Christof Stönner, Achim Edtbauer, Jonathan Williams

Max Planck-Institute for Chemistry, Mainz, Germany

Manuscript published in Indoor Air

Abstract Human beings emit many volatile organic compounds (VOCs) of both endogenous (internally produced) and exogenous (external source) origin. Here we present real world emission rates of volatile organic compounds from cinema audiences (50 - 230 people) as a function of time in multiple screenings of three films. The cinema location and film selection allowed high frequency measurement of human emitted VOCs within a room flushed at a known rate so that emissions rates could be calculated for both adults and children. Gas phase emission rates are analysed as a function of time of day, variability during the film, and age of viewer. The average emission rates of CO_2, acetone and isoprene were lower (by a factor of ~1.2 - 1.4) for children under twelve compared to adults while for acetaldehyde emission rates were equivalent. Molecules influenced by exogenous sources such as decamethylcyclopentasiloxanes and methanol tended to decrease over the course of day then rise for late evening screenings. These results represent average emission rates of people under real world conditions and can be used in indoor air quality assessments and building design. Averaging over a large number of people generates emission rates that are less susceptible to individual behaviours.

Keywords PTR-TOF-MS, Emission rate, Volatile organic compounds, Crowd breath, Movie theatre, Indoor air quality

Practical Implications The contribution of human emissions of volatile organic compounds (VOCs) to indoor air is an important yet often overlooked source of chemicals. The presented emission rates of VOCs averaged over a total of 8000 people can be used for characterizing indoor air influenced by human presence, source strength comparisons, building ventilation design and sick building syndrome assessments.

2.1. Introduction

Human beings are exposed to numerous volatile organic compounds (VOCs) in both outside and indoor air environments. More than half of the world's population now live in cities with significant airborne pollution[99] and exposure to outside air can have serious consequences for human health.[88] However, indoor sources of chemicals are also important, particularly since people spend much of their life (93% for the average American) in enclosed spaces such as buildings and vehicles.[79] Furthermore, as architects strive to improve the energy efficiency of buildings (e.g passive houses) the internal recirculation of air becomes key for heat conservation and hence indoor air quality becomes an important issue.[158] Known sources of indoor pollutants include building materials,[59, 90] carpeting,[60] furnishings[58] and products used or stored indoors such as paints[23, 86] and cleaning products.[105] Commonly reported indoor air pollutants include gases such as carbon monoxide, sulphur dioxide, nitrogen dioxide and ozone; microbial debris, selected VOCs and particulate matter.[11, 147] It has been noted that even when the emitted contaminants are present below threshold limit values,[158] they may contribute to a significant time-weighted exposure.[129] Humans too are a potent, yet often overlooked, source of chemicals to the indoor air environment. Several hundred VOCs have been reported emanating from human breath, saliva, skin, blood, milk, urine and faeces.[24] The major endogenous compounds emitted in human breath are acetone (1.2 - 1880 ppb), isoprene (12 - 580 ppb), ethanol (13 - 1000 ppb) and methanol (160 - 2000 ppb).[39] However, many other exogenous species may be uptaken (by inhalation and dermal uptake,[148] or on textile fabrics[141]) in outdoor polluted environments such as roadsides and subsequently re-emitted indoors, thereby effectively being imported into more confined domestic spaces. In this study, average VOC emission rates have been determined from a large number of people (8300) under real world conditions so as to include both endogenous and exogenous species. The aim is to provide a representative dataset of typical city dwelling human emission rates that can be used by architects, indoor air quality specialists, and medical researchers. Groups of people (50-238 at a time), were measured in a cinema which served as a convenient enclosed space that was ventilated at a known rate while the audience remained seated. By characterising the human emission rates of VOCs and CO_2 in the real world we may put other indoor sources into context and gauge the potential for indoor chemical reactions.
Here we present emission rates for numerous VOCs from seated human beings measured with a proton transfer reaction time-of-flight mass spectrometer (PTR-TOF-MS). This device allows quantification of numerous VOCs in real-time.[57] The measurements presented here took place in a cinema in Mainz (Germany) over a period of four weeks during the winter holidays 2015-2016. The study was designed to continuously measure from one screening room of the cinema. Physiological parameters or the exact age of the 8300 audience members were not recorded, although via ticketing information the proportion of the audience under 12 was known. Presented are the average VOC emission per person (above and below 12 years of age) from a crowd of people and how the main VOC emissions vary over time. The measurement of VOC emissions from a crowd neatly circumvents the tedious problem of sampling a statistically significant number of individ-

uals to encompass the main real-world source categories.[154] Such crowd measurement have been performed previously in enclosed, ventilated environments so that hundreds of people are monitored simultaneously, for example class rooms,[117, 135, 136, 144] office rooms,[126] other public buildings like a football stadium[142] and cinemas.[155] In contrast, much current breath research is focussed on the identification of biomarkers or chemical fingerprints from individuals to diagnose an illness.[6, 111, 115] However, the breath composition of an individual can vary significantly with dietary, sanitary and smoking habits, exposure to air pollutants,[41] position, and even the emotional state.[155] In future the results provided here may be compared with individuals to assess their representativeness and with disease biomarker candidates to gauge potential interferences.

2.2. Materials and Methods

2.2.1. Cinema/Movie Theatre

The measurements were made from screening room 2 in the Cinestar cinema complex in Mainz between 15 Dec 2015 and 15 Jan 2016. Within this time period, three films were screened: "Star Wars: The Force Awakens" and two German films "I'm off then" (German title "Ich bin dann mal weg") and "Help, I've shrunk my teacher" (German title "Hilfe, ich habe meine Lehrerin geschrumpft"). According to the "International Movie Database"[1] the "Star Wars" film falls under the genre "Action", "Adventure" and "Fantasy" whereas the other two films were "Comedy" films. The third film was additionally categorized under "Family" indicating that it was targeted at younger audiences. Star Wars was classified as suitable for viewing by people of 12 and above, while the other two films had no age restriction (USK 0). Nonetheless, the subject matter of "I'm off then" was more adult in nature (recounting a pilgrimage) whereas "Help, I've shrunk my teacher" was more directed at children. Screening room 2 has a capacity of 238 people and ticket sales, which are discounted for children (under 12 years old), permitted the proportion of children under 12 to adults to be known for each screening. The number of screenings, viewers and percentage of children (under 12 years old) in the audience are given in the supplement Table A.1. In total 8300 viewers were assessed over 85 separate film screenings.

The three films assessed in this study were screened at different times of day, from 11:30 in the morning to 22:30 at night. The summary of the screening hours of each film can be found in the supplement Table A.2. It is important to note, that the children's film "Help, I've shrunk my teacher" was screened mainly in the morning with only two screenings in the afternoon, whereas the other two films are distributed more evenly over the day. This distribution may bias the calculated VOC emission rates for this film and the consequences are explored later in the emission rates section.

The screening room was continuously flushed with outside air at a constant rate of 6500 m^3h^{-1}. All air was drawn in from the outside without any internal recirculation. Air entered the cinema through vents in the floor and was drawn out though opening

in the ceiling. The volume of the screening room was 1300 m^3 so that the overall air exchange rate was circa five times per hour. The entire exhaust airstream from the screening room was taken through a 75×75 cm stainless steel ventilation shaft to a separate technical room where the measurement instruments (PTR-TOF-MS and CO_2-Analyzer) were placed. In the middle of the ventilation shaft a 1/4" outer diameter (0.625 cm OD) Teflon sample line was inserted and 20 L/min air drawn continuously to the instruments.

2.2.2. Proton transfer reaction time-of-flight mass spectrometer

The VOCs in the screening room exhaust air were continuously monitored with a PTR-TOF-MS (proton transfer reaction time-of-flight mass spectrometer, PTR-TOF-MS 8000, Ionicon Analytik GmbH, Innsbruck, Austria). The ionization of the analyte molecules is made via hydroxonium ions (H_2O^+) which, due to the relatively low energies involved, results in little fragmentation of the analyte. The protonation occurs only for molecules possessing a higher proton affinity than water (691 kJ/mol), thus conveniently the system is blind to nitrogen, oxygen and argon, the main components of air.[12, 66] A detailed description of the set-up, the adjusted parameters of the PTR-TOF-MS and the calibration procedure is provided in the supplement.

2.2.3. CO_2 measurement

CO_2 was measured at 1 Hz using a commercially available Li-COR Li-7000 system. The linearity of the response was confirmed to 3400 ppmv using a second standard gas (10% CO_2, Air Liquide, Germany).

2.2.4. Data analysis

All modelling and statistical analyses were performed using the software R.[138] The data from each film title was divided into sections of the equal length corresponding to 30 minutes before the beginning of the film until 15 minutes after the end. Screening room background mixing ratios were sampled during the night between 3:00 and 6:00 local time when the cinema was closed to the public. In order to extract the masses that changed in the presence of the audience a paired Wilcoxon rank test for each mass was performed using the mean of the 15 minutes before the beginning of the film and the mean value during the film. A threshold p-value of 0.01 was chosen to extract the molecular mass signals which significantly increase in the presence of the audience.

2.2.5. Box model

Instantaneous emission rates were calculated by applying a mass-balance-approach. In this model it was assumed that there is no pathway for mass loss except air exchange and that the emission rate p is small compared to the air exchange rate. Since well-mixed conditions could not be assumed to apply in this model a correction variable

was introduced to account for the incompletely mixed air.[107] As stated previously, the volume of the screening room was 1300 m^3 and the air supply was 6500 m^3h^{-1} with identical flows in and out of the showroom.

$$\mathrm{dm/dt} = c_{in} \cdot q \cdot r \cdot + p - c_{out} \cdot q \cdot r \tag{2.1}$$

In Equation 2.1, m is the mass of the molecules at time t in the screening room air. The outside air is supplied with a ventilation rate r and a mixing ratio c_{in}. The mixing ratio c_{in} of each VOC in the inflowing air was interpolated from the two surrounding background night time measurements in the absence of people. The ventilation rate r is multiplied with the mixing parameter q, to account for the imperfect air mixing. The lower the mixing factor q the worse the mixing of air in the room providing a smaller effective room ventilation rate (product of $q \cdot r$). This accounts for the fact that the air less effectively mixed from the lower part of the cinema. This phenomenon leads to a slower exchange resulting in a flatter curve and in a later establishment of the steady state. The emission rate of a given gas from the audience is given by p. The effective ventilation rate was estimated by optimizing the mixing factor q and the emission rate of CO_2 p_{CO_2} with a normal least squares fit. The estimated mixing factor was used thereafter for all other calculations.
Rearrangement of Equation 2.1 results in an expression for the emission rate, Equation 2.2. For the calculation of the emission rate, the data was smoothed and differentiated using a Savitzky-Golay-filter with a span of 21 points and a polynomial order of 3.

$$p = \mathrm{dm/dt} - c_{in} \cdot q \cdot r + c_{out} \cdot q \cdot r \tag{2.2}$$

In Equation 2.2, the emission rate p expresses the total emitted mass per unit time [$\mu g\ h^{-1}p^{-1}$] for a specific number of viewers. Equation 2.3 was applied in order to determine the average emission rate per person (ER) while at rest. Henceforth, the emission rate (ER) will be referred to as the average emission rate per person.

$$ER = p/N \tag{2.3}$$

The mean of the total emitted mass p over the course of the entire film was divided by the number of viewers N and reported for each film. This results in an emission rate of grams (of a particular molecule) per hour per person. All calculated values for the emission rate are presented as the average emission rate per person [$\mu g\ h^{-1}p^{-1}$]. The cinema provided also the number of children (younger than 12 years) watching the film based on the ticket sales. The mean emission rates per person estimated from the "I'm off then"-film (ER_{adults}) were used to calculate the emission rate of children. To do this it was assumed that the emission rate for children during the child film "Help, I've shrunk my teacher" is the difference between the total emission rate p minus the sum of the emission rate of the adults ($\sum ER_{\mathrm{adults}}$).

$$p_{\mathrm{children}} = p - \sum ER_{\mathrm{adults}} \tag{2.4}$$

The division of the total amount of emitted VOC per minute by the amount of children (Equation 2.3) resulted in the emission rate per child. The "I'm off then" -film was only attended by adults because of the subject matter (a pilgrimage), even though the film is free for all age groups (unrestricted), see Table A.2. Thus the emission rates calculated from these screenings were labelled as pure "adult". In the case of the "Star Wars"-film the audience consisted of people from different age classes, beginning at the age of 12. The film "Star Wars" is more directed at younger viewers than the film "I'm off then" and therefore the emission rate obtained from the Star Wars screenings were labelled as "mixed". Given the emission rate of "adults" calculated from the "I'm off then" screenings the emission rates labelled as "children" were obtained by applying Equation 2.4.

2.3. Results and Discussion

2.3.1. Calculated effective ventilation rate and results of the box model

The mean mixing factor derived from the CO_2 data from all screenings was found to be 0.3 ± 0.1 with a residual sum of squares ranging from 0.97 to 0.99. Previously reported literature values for this parameter range from 0.1 for imperfectly mixed rooms to 1 indicating fully mixed.[107] There is no dependence of the mixing factor on the number of viewers (correlation coefficient r = -0.07).
Further calculations were performed using the mean of the mixing factor giving q= 0.3. The second parameter estimated by the model is the emission rate of CO_2 per person. The calculated emission rate p_{CO_2} was estimated to be $2.9 \cdot 10^7$ µg $h^{-1}p^{-1}$ with a standard deviation of $0.1 \cdot 10^7$ µg $h^{-1}p^{-1}$.

2.3.2. Emission rates

In Figure 2.1, the CO_2 mixing ratio (black) and the emission rate per hour per person (grey) are shown. Most of the measured VOC from human beings show a similar general behaviour during all screenings with a slow steady increase as the audience enters the previously empty screening room and a steep decrease at the end of each film, when the audience departs. Typical mixing ratios of CO_2 lie between 400 and 2500 ppm, for acetone between 3.00 and 20.00 ppb, and for isoprene between 1.00 and 9.00 ppb. These molecules are known to be endogenously produced and emitted in human breath.[7, 24] The emission rate at the beginning of each film cannot be evaluated quantitatively since the audience enters the screening room little by little and the door is open to the foyer area. During the film, the elevated emission rate remains on average reasonably stable albeit with clearly defined peaks at certain moments (associated with events in the film), and then decreases rapidly at the end of the film. The mean emission rates for the measured VOCs were calculated from the beginning of the film until 5 minutes before the end.

The mixing ratio of the VOCs entering the screening room (c_{in}) was estimated using

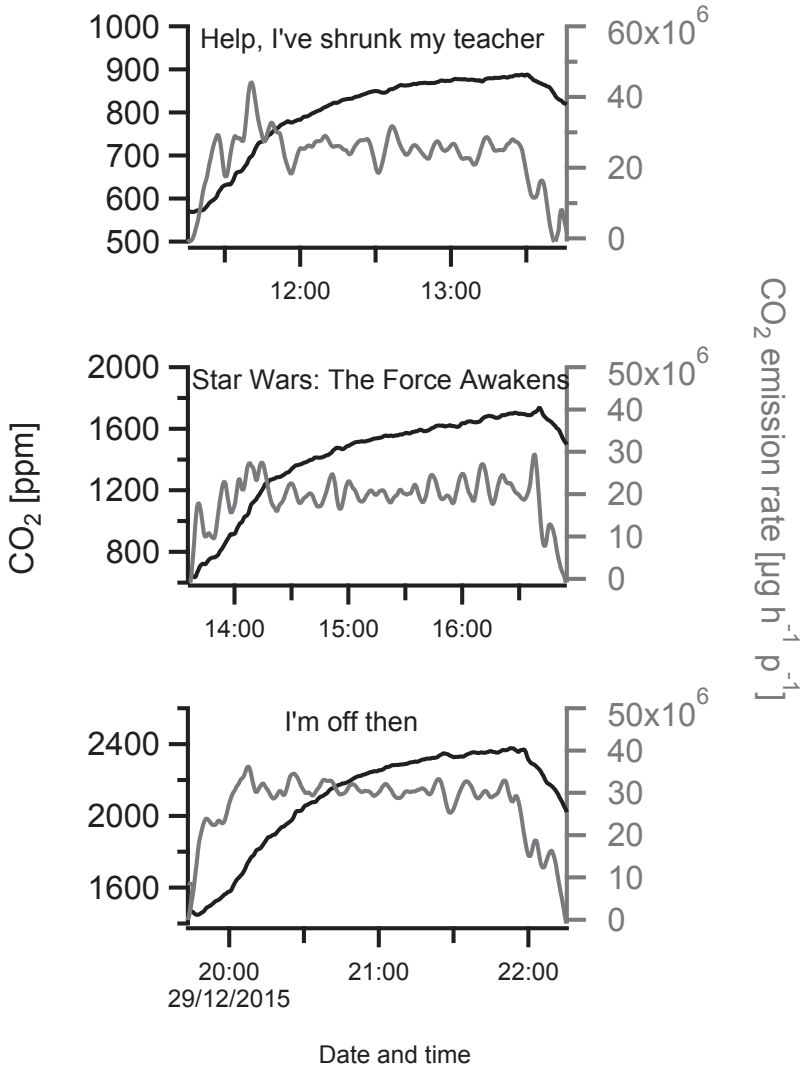


Figure 2.1.: Behaviour of the emission rate per person (grey) and the mixing ratio (black) of CO_2. The top panel shows the film "Help, I've shrunk my teacher", the middle panel "Star Wars: The Force Awakens" and the bottom panel "I'm off then".

the interpolated value between the two night-time background measurements. Assuming lower mixing ratios of VOCs during the day time (true for CO_2) than during night time background measurements, the emission rate would be higher than reported. For CO_2 a maximum error of 30% was calculated for 54 viewers and a mixing ratio of CO_2 of 627 ppm at night. This condition depicts the maximum error and mixing ratios higher than 600 ppm at night were measured only for 5 films. The average mixing ratio during night was 483 ± 44 ppm CO_2 resulting in an error of 15% with 54 people attending the cinema. The higher the amount of viewers attending the cinema the smaller the error

becomes (on average of 98 people attended the screenings). In general, we assume that the error introduced in the emission rates by different diurnal VOC concentrations is small compared to the standard deviation presented in Table 2.1.
The emission rates of selected gaseous species from children and adults, provided by Equation 2.3 and 2.4, are shown in Table 2.1. The presented masses were either calibrated with the use of a gas standard or calibrated using the calibration factor of acetone (30.9 ncps/ppb). An overview of all detected VOC signals can be found in the supplement Table A.3.

Table 2.1.: Emission rates of various VOC and CO_2.

Molecule	Pr. Mass [m/z]	Adults [$\mu g\, h^{-1} p^{-1}$]	Std.dev. [$\mu g\, h^{-1} p^{-1}$]	Children [$\mu g\, h^{-1} p^{-1}$]	Std.dev. [$\mu g\, h^{-1} p^{-1}$]	Cal. method
Carbon dioxide	-	$3.0 \cdot 10^7$	$0.5 \cdot 10^7$	$1.8 \cdot 10^7$	$0.6 \cdot 10^7$	Cal. gas
Formaldehyde	31.0178	207	104	426	375	Cal. factor
Methanol	33.0335	650	736	1136	984	Cal. gas
Acetaldehyde	45.0335	221	76	252	160	Cal. gas
Ethanol	47.0491	216	154	116	171	Cal. factor
Acetone	59.0491	419	96	333	202	Cal. gas
Isoprene	69.0699	166	39	95	59	Cal. gas
Sum of all monoterpenes	137.1325 + 81.0699	201	170	189	181	Cal. gas
Siloxane (D_5)	355.0698	112	104	256	186	Cal. factor

In order to distinguish between exogenous and endogenous emissions, the standard deviation of the emission rates was examined over several screenings of the same film. We hypothesize that exogenous sources will be significantly more variable over the time of day. This is supported by the relatively small standard deviations that were observed for CO_2, acetaldehyde, acetone and isoprene which are known to be predominately endogenous. The protonated mass of decamethylcyclopentasiloxane (D_5) would be m/z 371.0956 but the most abundant peak appears at m/z 355.0698 due to the elimination of a methyl group. Based on their relatively high standard deviation, we consider ethanol, the siloxane, methanol and monoterpenes to be predominately exogenous.
The emission rate per person for CO_2 was estimated to be 30 ± 5 g $h^{-1}p^{-1}$ for adults and 18 ± 6 g $h^{-1}p^{-1}$ for children. Persily et al.[110] derived the CO_2 emission rates from well-established concepts concerning the human metabolism and physical activity. Assuming an average age between 21 to < 30 and a physical activity of 1.4 met (between 1.3 for "sitting, reading, writing, typing" and 1.5 for "sitting at sporting event as spectator". The unit "met" quantifies the level of physical activity) we calculate an emission rate of 40 g $h^{-1}p^{-1}$ for males and 32 g $h^{-1}p^{-1}$ for females. The emission rate decreases continuously for younger or older people (both male and female). For children (younger than 12) the reported value of 18 ± 6 g $h^{-1}p^{-1}$ underestimate the emission rate

calculated by Persily et al. for an age class between 6 to < 11 lying between 22 to 25 g $h^{-1}p^{-1}$ for males and 19 to 23 g $h^{-1}p^{-1}$ for females.
Tang et al.[136] recently published emission rates from several VOCs measured in a classroom. Table 2.2 compares the emission rates calculated by Tang et al. with the values presented in this study. Especially, the emission rates per person of isoprene, monoterpenes and the (iso)butyl fragment are in good agreement. For ethanol the adult emission rate from the cinema is higher (216 µg $h^{-1}p^{-1}$ for adults) which comes from the consumption of alcoholic beverages in the evening also resulting in a higher standard deviation. Comparing the ethanol emission between pre-evening and evening screenings the estimated emission rate is calculated to be 132 µg $h^{-1}p^{-1}$ for the pre-evening screenings (before 18:00) and 329 µg $h^{-1}p^{-1}$ for the evening screenings (18:00 and later, compared to 94.9 µg $h^{-1}p^{-1}$ from Tang et al.). Tang et al. summarized all sulfur-containing compounds resulting in an emission rate of 6.5 µg $h^{-1}p^{-1}$ which is close to the emission rate of dimethyl sulfide or ethyl mercaptan derived in this study. The emission of methanol is discussed later in more detail but was found to be variable over the day exhibiting high values in the morning.
The skin oxidized VOC like 6-methyl-5-heptene-2-one (6-MHO) or 4-oxopentanal (4-OPA) were less abundant or were not detected in the cinema. Possible explanations might be the lower ozone mixing ratios in winter or the effective removal of ozone within the intake of the ventilation system and hence a lower amount of oxidation products. A previous study measuring the air within an open air football stadium using the same instruments in summer reported a signal of 6-MHO.[157] Another explanation could be that these products were already evaporated from the skin during the waiting time in the foyer of the cinema which would have low ambient ozone due to effective indoor deposition. This could be also true for acetone which is reported to be a product of skin lipid ozonolysis, too.[127] Compared to the emission rate presented in this paper of 419 µg $h^{-1}p^{-1}$ Tang et al. reported a value twice as high. In 1975 Wang et al.[144] published a study concerning emission rates of bioeffluents from humans. In general, their calculated emission rates lie above those presented in our study with the exception of CO_2. However, bearing in mind that the study from Wang et al. was conducted 40 years ago that most of the values are in the same order of magnitude is reassuring.

Table 2.2.: Summarization of emissions rate of several VOC from this study and Tang et al.[136]

Molecule	Protonated Mass [m/z]	Tang et al. [µg $h^{-1}p^{-1}$]	Adults [µg $h^{-1}p^{-1}$]	Children [µg $h^{-1}p^{-1}$]
Acetone	59.0491	1060	419	333
Acetic Acid	61.0284	329	205	357
Methanol	33.0335	156	650	1136
Acetaldehyde	45.0335	114	221	252
Monoterpenes	137.1325 + 81.0699	187	201	189
Isoprene	69.0699	162	166	95
Ethanol	47.0491	94.9	216	116
$C_6H_{10}H^+$	83.0855	88.8	22	32
(iso)butyl fragment	57.0699	39.7	41	52
Propionic acid / hydroxyacetone	75.044	40.4	19	27
6-MHO	127.1168	99.3	3	5
(iso)propyl fragment	43.0542	23.8	107	321
S-containing	63.0263	6.5	7	6

Figure 2.2 shows the emission rate in µg per hour per person for different VOCs in a boxplot. The black solid line in the box indicates the median and the boxes encompass the 25 and 75 percentiles of the data. The whiskers are 1.5 times the interquartile range. The different colours indicate the age classification of the film. The molecules shown in Figure 2.2 were those with the highest emission rates measured, whereby CO_2 was by far the greatest emission source (ca.30 g $h^{-1}p^{-1}$) followed by methanol and acetone (approximately four orders of magnitude less) along with the other VOCs.

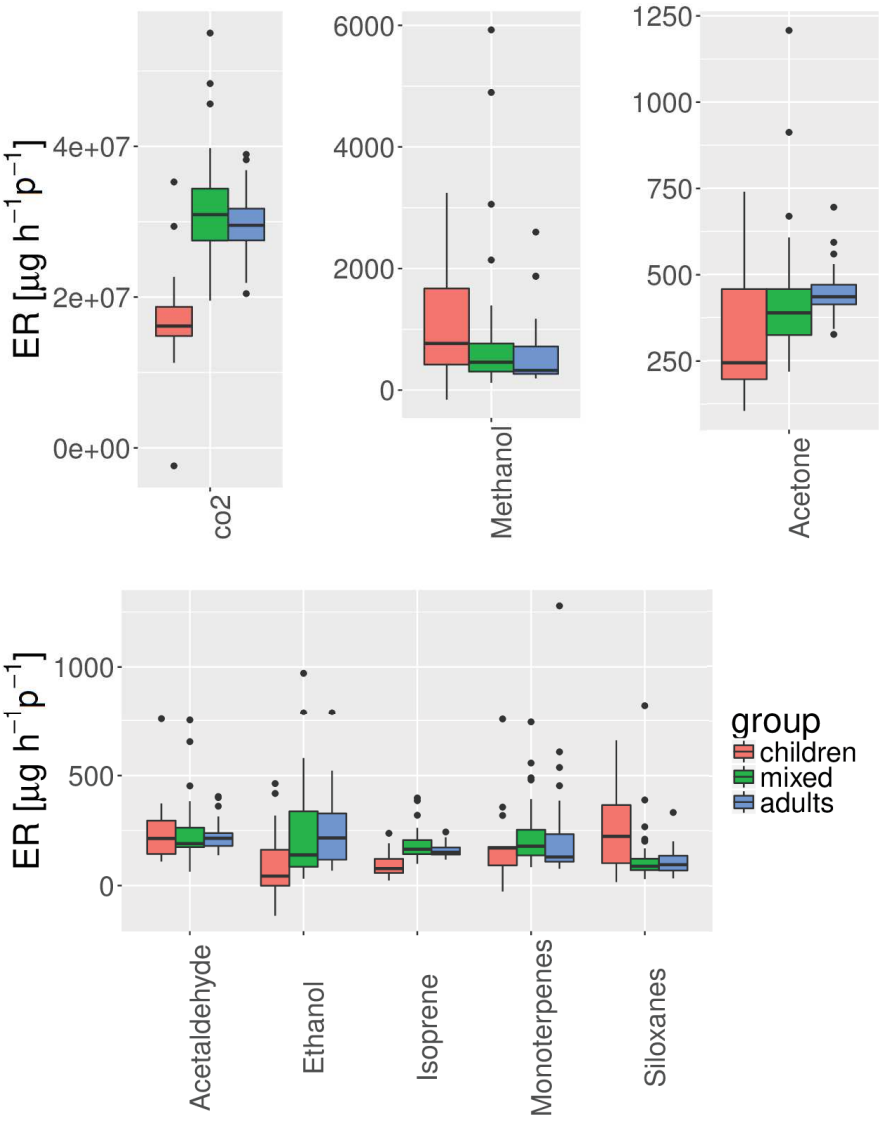


Figure 2.2.: Boxplots for different VOCs and different age groups. In the upper part carbon dioxide, methanol and acetone are shown (from left to right). The lower part includes acetaldehyde, ethanol, isoprene, pinene, monoterpenes and decamethylcyclopentasiloxane.

Some of the VOCs show significantly different emission rates depending on the age classes. CO_2, acetone and isoprene show similar behaviour with the lowest emission rates for children and highest for adults. For CO_2 and isoprene, the emission rates for the mixed audience ("Star Wars", rating 12) and the adults only ("I'm off then", rating 0) are almost equal but they differ significantly from the emission rates of children (calculated as shown in Equation 2.4). Figure 2.2 shows a slightly higher emission rate for CO_2 for the mixed group than for adults. An explanation could be that the "Star Wars"-film from which the emission rates of the mixed group were derived were screened mostly at 14:00 (20 screenings) and 22:30 (12 screenings). In the case of the screening at 22:30 only people of an age of 18 or older are allowed to enter and the target audience of the "Star Wars"-film is probably younger than that of the "I'm off then"-film (recounting a pilgrimage). This group specifically people between 21 to < 31 emit the highest amount of CO_2. The calculated emission rate for CO_2 at 22:30 screening the "Star Wars"-film is 36 ± 4 g $h^{-1}p^{-1}$ compared to 28 ± 5 g $h^{-1}p^{-1}$ at 14:00 (see Figure 2.3). There is no data available on the average age or gender distribution for the people attending the cinema. The age dependency of the isoprene emission rate is in agreement with the previously reported age dependency of isoprene in human breath described in Lechner et al.[87] Therein it was recognised that children emit significantly less isoprene than adults. As for CO_2 the emission rate for isoprene is slightly higher for the mixed group than for the adults also showing an enhanced emission rate at 22:30 (233 ± 58 µg $h^{-1}p^{-1}$ at 22:33 and 148 ± 23 µg $h^{-1}p^{-1}$ at 14:00). Lechner et al.[87] reported a significant lower isoprene emission rate for 19-29-year-old subjects than for older adults. However, it should be noted that only 11 subjects were measured in 19-29-year-old age category in the Lechner et al. study. A much earlier study by Mendis et al.[95] also reported no age dependency of isoprene concentration in expired air, sampling from 43 healthy volunteers between 22 and 75 years. It should be noted that this study differentiates only between children up to 12 years and older persons, based on ticket sale information. Thus the exact average age of the viewers attending the measured films could not be determined. Interestingly, the function and source mechanism of isoprene is still a matter of debate. Isoprene is clearly endogenously produced and it is suggested that its production is linked to the cholesterologenesis.[131]
The emission rate for acetone shows significantly different emission rates between the children and adult age classes with a p-value of 0.05. A similar difference in breath acetone, whereby children emit less than adults, has been also reported by Enderby et al.[35] In our study the "mixed" age class emission rate lies in between children and adults. However, Enderby et al found no correlation between breath concentration and age in an age range between 7 to 18 years.[35] Acetone is produced by the liver during fatty acid metabolism which acts when glucose energy sources are not available. Higher levels of acetone in blood are therefore measured in humans during fasting and prolonged exercise.[84] The larger acetone emission rates from adults (13% higher) may be simply a function of their larger body mass. Further factors like the diabetic status of the attendees presumably leading to higher emissions rates of acetone cannot be excluded.[145]
Acetaldehyde was found to be age independent, so emission rates of children and adults are comparable despite differences in body mass. This is in agreement with previous

studies.[35, 140], Acetaldehyde is produced in the liver as an intermediate in the ethanol metabolism[42] and through the action of bacteria on ethanol in the mouth.[143] The aforementioned molecules all show a relatively small standard deviation compared to their emission rate and to the other compounds. This behaviour may reflect the fact that they are predominantly endogenously produced and thus are less liable to influence by exogenous factors like food and drink consumption.
The average CO_2 emission rate for the different films and screening times can be seen in Figure 2.3. The CO_2 emission rate during each of the films ("Help, I've shrunk my teacher" in red screened at 11:30 and 17:20, "Star Wars" in green screened at 14:30,18:00 and 22:30 and "I'm off then" in blue screened at 17:30 and 20:00) are closely comparable for multiple screenings of the same film, but the three films exhibited significantly different emission rates and standard deviations, see Table 2.1. Interestingly the emission rate for CO_2, as well as for acetone and isoprene (and many other species) gets higher during later screenings with a maximum at 22:30. This might be a result of a higher emission rate seen in the example of CO_2 or the underestimation of the emission rates during midday due to the error introduced by the measurement of the background during night. The CO_2 emission rate for the children's film, shown at 11:30 and 17:20 (only two screening times), lay well below that of the other two films.
When examining the data for trends in the emission rate as a function of time of day, it is important to note that all "Help, I've shrunk my teacher" films were screened in the morning and early afternoon whereas the "Star Wars" and "I'm off then" films had screening times distributed over the day, as shown in the supplement Table A.2. In Table 2.1 it can be seen that other VOCs like methanol, ethanol and the monoterpenes show larger standard deviations compared to their emission rates than the main breath gases discussed above. Methanol is known to be produced endogenously by the consumption of fruit through the degradation of pectin.[91] The high standard deviation may stem from the different dietary habits between the viewers. Therefore, the high emission rate of methanol for the children's film "Help, I've shrunk my teacher" may be caused by the consumption of fruits and fruit juices during breakfast since this emission rate diminishes during the day. The middle panel in Figure 2.3 depicts the daily pattern of methanol which is clearly distinct from CO_2. The emission rate of methanol shows a maximum at 11:30 and again at 22:30 and a relatively constant emission rate during the rest of the day. Monoterpenes are ingredients of many fruits and can be also found in soft drinks like Cola. Additionally, monoterpenes like limonene are frequently used as fragrances in personal care and cleaning products. The use of personal care products and their effect on emission rates is discussed in detail using the example of decamethylcyclopentasiloxane below.
In the lower panel the emission rates of decamethylcyclopentasiloxane (D_5) calculated at different screening hours is shown. This molecule belongs to a group of chemicals collectively called siloxanes or silicones that are commonly found in personal care products such as shampoo and deodorants as well as in cleaning and polishing products.[62] It is therefore an exogenous species and shows a temporal emission behaviour similar to that reported by Tang et al. who measured the mixing ratio of different cyclic siloxanes in a classroom of engineering students.[135] In that study D_5 was found to decrease over the

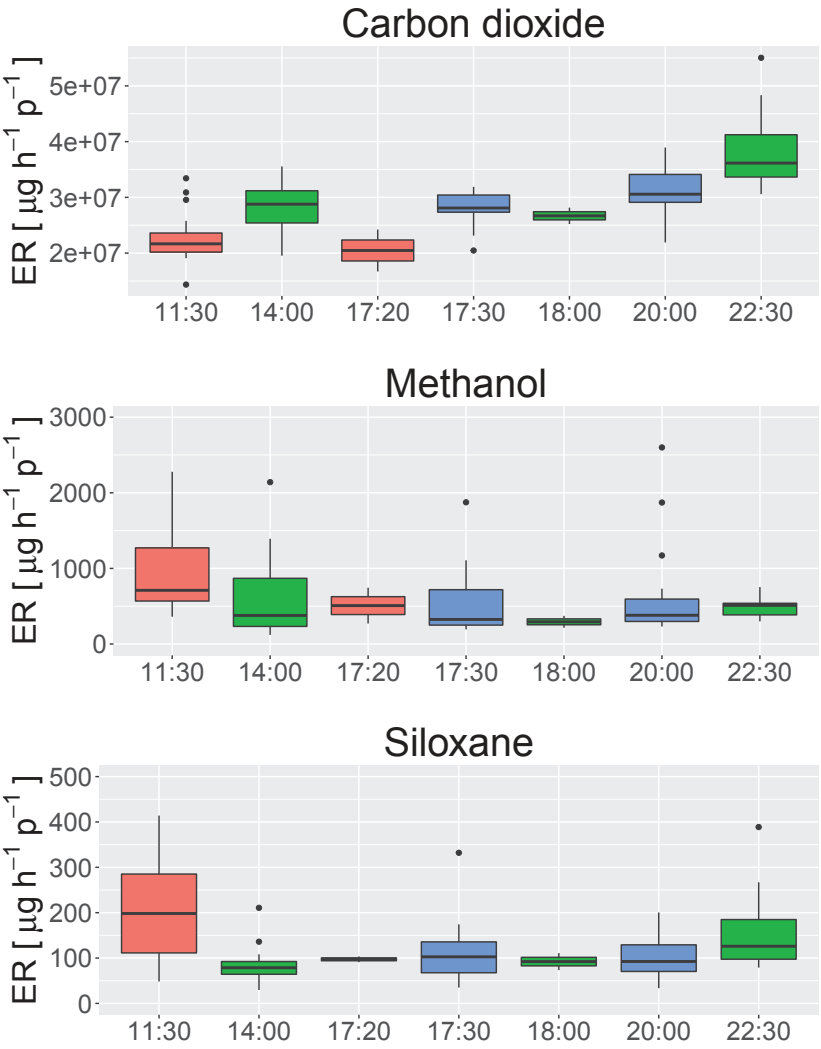


Figure 2.3.: Emission rates of CO_2 (top) and methanol (middle) and decamethylcyclopentasiloxane (bottom) during the course of the day. The colours indicate the film screened in the showroom. The film "Help I've shrunk my teacher" is shown in red, "Star Wars" in green and "I'm off then" in blue.

course of day, likely due to evaporative losses of applied hygiene products. Indeed, the earliest and the latest screenings show the highest values. This may simply reflect that hygiene and cosmetic products containing siloxanes are applied in the morning then "refreshed" later on prior to late screenings. The emission rate of several different siloxanes were calculated in a study of Tang et al.[135] The calculated emission rates provided by Tang et al. were generally higher in the morning than in the evening due to outgassing from siloxanes of cosmetic products, ranging for D_5 from 9800 to 183 µg $h^{-1}p^{-1}$. In our study such strong differences for D_5 could not be found, as shown in Figure 2.3, most of our reported average values lay close to the afternoon levels (between 14:10 and

16:00 pm) of 183 µg $h^{-1}p^{-1}$. This is despite the ventilation rates in both rooms being comparable. This discrepancy may reflect the fact that we used the average over the film to calculate VOC emission rates, whereas Tang et al. reported emission rates every minute. This is important because higher emission rates for D_5 were observed at the beginning and at the end of the film, when the audience undress and dress respectively. This behaviour is shown in Figure 2.4 along with the emission rate of CO_2. The peak emission of D_5 for the films screened in the morning was on average 2800 µg $h^{-1}p^{-1}$, and was therefore only slightly higher than measured for the films in the afternoon with an emission rate of 2500 µg $h^{-1}p^{-1}$. In general, we used only values during the film and not the peak emission at the beginning of the film since as the audience is entering the cinema we do not know the exact number of people present.

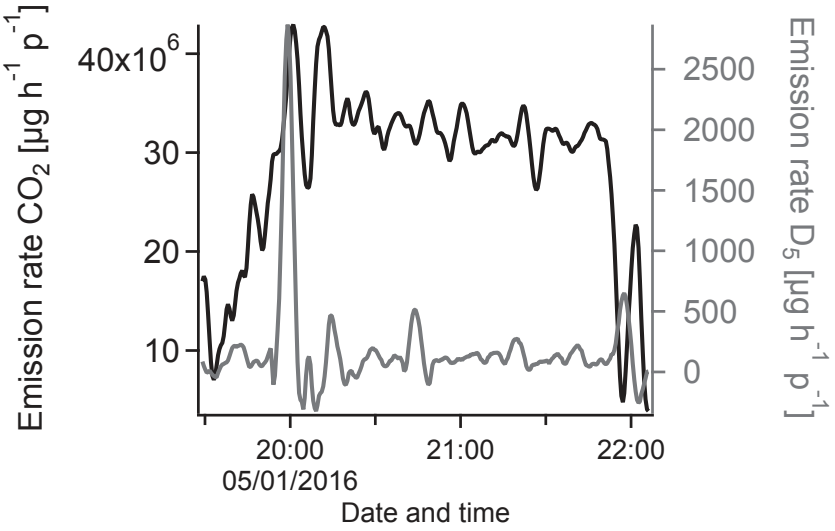


Figure 2.4.: Emission rate of CO_2 (black curve) and Decamethylcyclopentasiloxane (grey curve) in [µg $h^{-1}p^{-1}$] during the film "Star Wars" starting at 14:00 and ending around 16:30.

In Figure 2.4 the emission rate of D_5 decreases during the film while the audience remains seated since it is not emitted from breath and is probably emitted from skin at lower rates when the audience does not move. This indicates that they are not released from the human metabolism but from a “burst” source associated with ruffling of hair, clothes and skin.[135] All other VOCs exhibit some emission source during the film that varies with time. The examples of D_5 show clear emission rate changes over the day. Determining differences between children and adults in our study is complicated by these temporal trends and by the fact that family films are screened earlier. The time of day can be an important factor for the release of VOCs due to temporally dependent habits and metabolic differences. Evaporative loss influences emission rates of chemicals which are applied or consumed only several times a day.

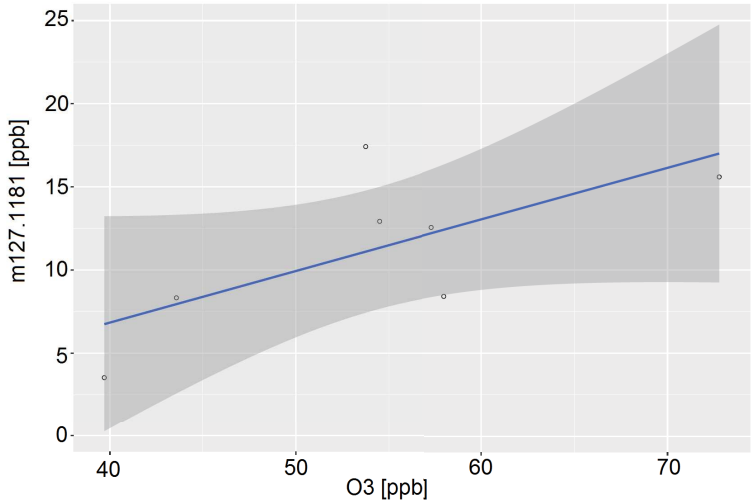
2.4. Conclusion

In summary, a PTR-TOF-MS and a CO_2 instrument were plugged into the exhaust ventilation shaft of a movie theatre in order to characterize the emissions of VOC and CO_2 from a large number of seated people under real world conditions. By sampling a crowd of people at different times of day rather than individuals a representative average human VOC emission is obtained over a broad range of dietary and smoking habits, activity level, state of health, environmental exposures, age, stress level or mood. This approach offers a statistically robust method for determining average emission rates of VOC from humans incorporating multiple sources including breath, skin, clothes and some foodstuffs.[154]
The most abundant compounds were the endogenous breath compounds CO_2, acetone and acetaldehyde, and the predominantly exogenous ethanol, monoterpenes and decamethylcyclopentasiloxane (D_5). The emission rates of the VOCs measured from humans covered a range of 5 orders of magnitude, and CO_2 emission rates were a factor of 105 higher than the VOC emissions. Large variances were found between adults and children younger than 12 (for CO_2, acetone and isoprene), for time of day (methanol, siloxanes),[39, 135] and between seated and moving crowds. The underlying reasons for the differences can be biological (result of metabolic processes as with isoprene and acetone) or behavioural (from hygiene or diet as with siloxanes and methanol). Since this dataset represents average emissions from a wide cross section of society it can therefore be used for indoor air chemistry studies, comparison of source strengths, and building design.

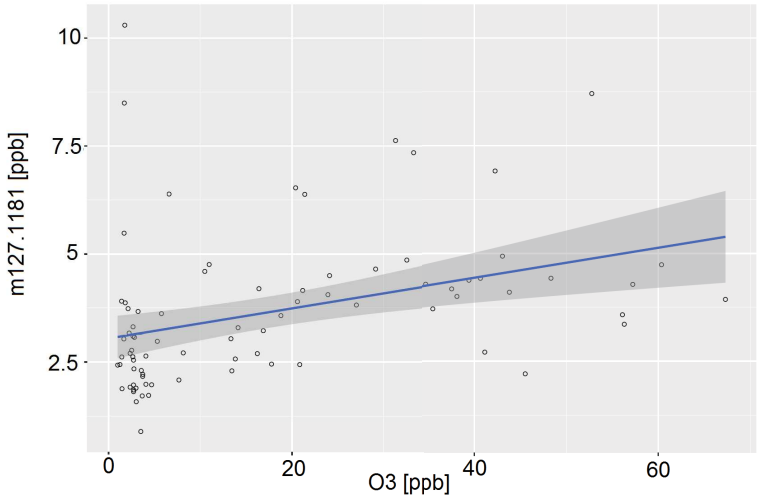
3. Investigation of the emission of VOCs from humans as a function of the ambient ozone mixing ratio

The previous section shows the emission rates of VOCs emitted by humans. These can be classified in: 1) mainly endogenously produced such as CO_2, acetone and isoprene and 2) exogenously produced such as ethanol and decamethylpentasiloxane. It was found that the emission rates depend on biological and behavioural differences. Recent work showed the production of oxidated species due to the ozonolysis of skin products. It was shown that 6-methyl-5-hepten-2-one (6-MHO) is one of the most dominant degradation products.[142, 149, 156, 157] This compound was also measured in the cinema showroom in both winter and summer. It is assumed that the largest part is transported from the outside into the screening room since ozone is effectively removed on surfaces such as metal ventilation shafts and air filters and thus negligible ozonolysis is assumed to occur the showroom.
The measurement of VOCs in a showroom of a cinema took place in summer 2016 for four weeks during the European Football Championship. Additionally, ozone data was retrieved from the Mainz Umweltbehörde from a measurement station in Mainz-Mombach. The ozone data does not come exactly next to the cinema. The measurement station is located 6 kilometres to the south from the cinema. However, we assume that the measurement represents the regional ozone level and can be taken to show seasonal differences. We do not perform any quantitatively analysis with this ozone data. The ozone mixing ratio was calculated by averaging over 3 hours before the beginning of the film.

Figure 3.1 shows the relationship between 6-MHO and ozone for summer 2016 and winter 2014/2015. The mixing ratio of 6-MHO in the cinema ranges from 3.5 ppb to 17.4 ppb with a median value of 12.6 ppb during summer and from 0.9 ppb to 10.3 ppb with a median value of 3.3 ppb during winter. The ozone mixing ratio ranges from 40 ppb to 73 ppb with a median value of 55 ppb in summer and from 1 ppb to 67 ppb with a median value of 15 ppb in winter. The correlation coefficient is $r = 0.70$ in summer and $r = 0.36$ in winter. This supports the suggestion that 6-MHO is formed via surface ozonolysis of skin oils. Due to the small data set size during summer no further investigation was performed.
The comparison of average emission rates between summer and winter revealed higher emission rates for 6-MHO during summer (~3.8 times higher in summer) whereas decamethylsiloxanes (~1.6 times higher in winter) and monoterpenes (~2.5 times higher in winter) were emitted in higher amounts in winter. Possible explanations might be



(a) Summer



(b) Winter

Figure 3.1.: Scatter plot between 6-MHO and ozone during summer (upper panel) and winter (lower panel). The blue line shows the estimated linear regression line with its error as the dark grey shaded area.

higher emission rates of these exogenous compounds in winter due to different seasonal hygiene habits. Another reason might be the difference in outside temperature and thus the higher evaporation of these compounds in summer. Indeed, the meteorological conditions (such as ozone mixing ratio, temperature and rainfall) can have an impact on the release of VOCs in indoor environments and should be taken into consideration. For 6-MHO a positive correlation ($r = 0.40$) was also found with temperature (with a correlation coefficient of $r = 0.64$ between temperature and ozone) and statistical methods might provide useful tools for discovering these dependencies.

4. European football: Goals change crowd air chemistry

Christof Stönner and Jonathan Williams

Max Planck-Institute for Chemistry, Mainz, Germany

Manuscript published in Nature Correspondence

During live public screenings of the 2016 UEFA European Championships, the emission rates of particular chemicals in the audience's breath vary sharply - apparently in response to events on the football pitch.

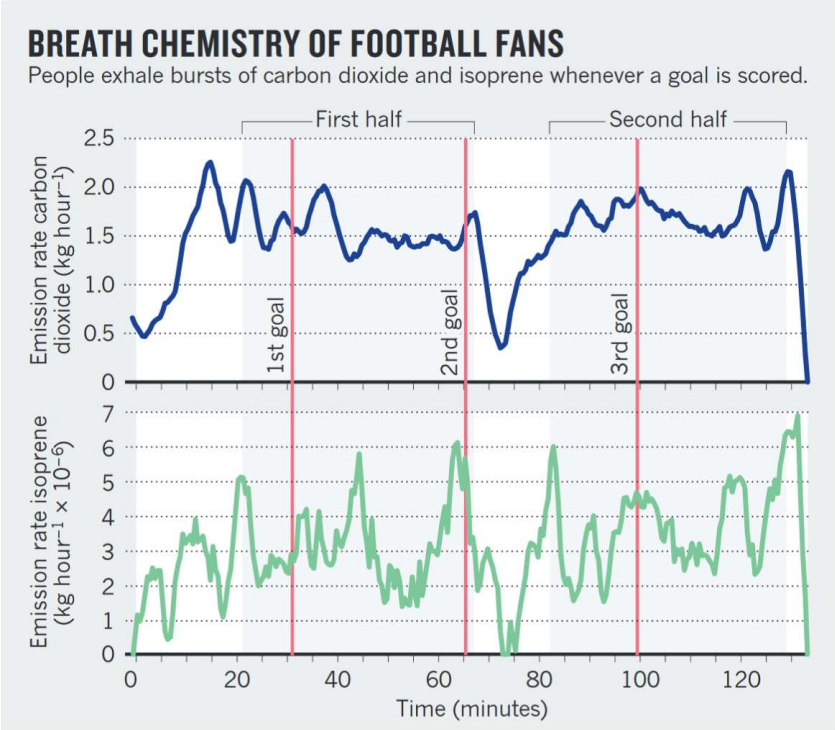


Figure 4.1.: People exhale burst of carbon dioxide and isoprene whenever a goal is scored.

Football matches induce fans to roar in jubilation, hold their breath in suspense and sigh with disappointment. On 26th June, we tracked reactions from a cinema audience

during the Germany - Slovakia game by monitoring changes in air composition resulting from their exhalations (for methodology, see Williams et al.[155]).
In moments of high excitement, exhaled carbon dioxide seems to spike as people's heartbeats and breathing accelerate (see 'Breath chemistry of football fans'). So do emission rates of isoprene, which is released from muscles as fans spring from their seats when a goal is scored. Breath chemistry therefore appears to ride the same emotional roller coaster as the live broadcast.

5. Cinema audiences reproducibly vary the chemical composition of air during films, by broadcasting scene specific emissions on breath

Jonathan Williams[1], Christof Stönner[1], Jörg Wicker[2], Nicolas Krauter[2], Bettina Derstroff[1], Efstratios Bourtsoukidis[1], Thomas Klüpfel[1], Stefan Kramer[2]

[1] Max Planck Institute for Chemistry, Mainz, Germany
[2] Johannes Gutenberg University of Mainz, Germany

Manuscript published in Scientific Reports

Keywords Atmospheric chemistry Computer science Diagnostic markers Human behaviour

Abstract Human beings continuously emit chemicals into the air by breath and through the skin. In order to determine whether these emissions vary predictably in response to audiovisual stimuli, we have continuously monitored carbon dioxide and over one hundred volatile organic compounds in a cinema. It was found that many airborne chemicals in cinema air varied distinctively and reproducibly with time for a particular film, even in different screenings to different audiences. Application of scene labels and advanced data mining methods revealed that specific film events, namely "suspense" or "comedy" caused audiences to change their emission of specific chemicals. These event-type synchronous, broadcasted human chemosignals open the possibility for objective and non-invasive assessment of a human group response to stimuli by continuous measurement of chemicals in air. Such methods can be applied to research fields such as psychology and biology, and be valuable to industries such as film making and advertising.

5.1. Introduction

All living organisms from the smallest plants and bacteria to trees and primates emit chemicals into their local environment.[24, 56, 75, 159] Such chemicals may act as signals, eliciting wide ranging responses[31, 70]. The atmosphere has been shown to be

an effective conduit for chemical communication between plants and plants,[10] plants and insects,[121] insects and insects.[5] Yet the extent, or even existence of airborne chemical communication between humans remains controversial.[32, 162] Despite reported chemosignal volatiles in human tears affecting testosterone levels,[45] armpit and sweat odours interpreted as fear signals,[2, 4, 29] sleeping babies responding to lactating breast volatiles[33, 93, 119] and menstrual synchronization,[130] no human pheromone (an evolved chemical signal between humans) has been reliably and reproducibly identified.[160] Generally, studies reported to date have been small in scale (number of people and measurements), subjectively assessed,[2, 29] and often with unnaturally high concentrations of bioassays, due to the analytical methods available. To screen groups of people for potential emotion signalling molecules at natural levels we have conducted a largescale study involving more than 9500 cinemagoers who viewed 108 screenings of 16 different films (including comedy, horror and romance, see Table 5.1). During the films, audiences were subjected to audiovisual stimuli while outside air was directed into the cinema through floor vents and out through ceiling vents (normal operating practice), and in the outflow, the concentration of over 100 trace gas species was measured using proton transfer reaction time-of-flight mass spectrometer (PTR-TOF-MS) and infra-red spectroscopy. Data was collected at 30 second time resolution and with sub-ppb(10^{-9}) detection limits to investigate potential causal links between the audiovisual stimuli and audience emitted chemicals.

Table 5.1.: Summary

Film	Number of Screenings
Buddy	17
Walking with Dinosaurs	15
The Hobbit - The desolation of Smaug	15
The Secret Life of Walter Mitty	15
The Hunger Games 2	10
Carrie	7
Suck me Shakespeare	5
The Little Ghost	4
Journey to the Christmas Star	4
Paranormal Activity 6	4
Belle and Sebastian	3
The Counselor	3
Machete Kills	2
Cloudy with a chance of Meatballs 2	2
The Physician	1
Bolshoi: Sleeping Beauty	1
Total	108

Of the 872 volatile compounds identified in human breath[24], a fraction is thought to be produced endogenously. These compounds can be used to track chemical changes within the body, over long (with age)[83, 87] and short timescales (medication response, food, disease or exercise).[75, 76, 128, 131] Within this cinema based study we hypothesize that if films elicit strong emotional responses then volatile products from the internal biochemical response (cardiovascular, skeletomuscular, neuroendocrine, and autonomic nervous system)[34, 89] may be vented shortly afterwards over the lungs, and observed as transient peaks in concentration in air exiting the cinema. Full details of the experimental set-up and instrumentation is given in the method section.

5.2. Results

Figure 5.1 shows sections of the CO_2 data measured in air from the Mainz Cinestar cinema. In Figure 5.1a, large CO_2 peaks can be observed between 26th and 30th December, each corresponding to the screening of a particular film. Prior to a film starting in the empty cinema, CO_2 approximates to background levels (ca. 400 ppm) as ambient air is continually drawn through the cinema from outside. People exhale air with circa 4% CO_2, so that as the audience arrives, CO_2 levels increase, rapidly at first, and then more slowly as the equilibrium value is approached after about ninety minutes, reaching levels between 1000–2400 ppm. This is some 2 to 8 times the current ambient background levels (400 ppm), but well below the European indoor standard limit of 3500 ppm. In effect, the cinema is a small scale analogue of the on-going planetary scale increases in CO_2 in which additional anthropogenic CO_2 sources from fossil fuel usage must equilibrate with the slow uptake rates into the ocean, vegetation and soils.[40] At the end of each film the CO_2 level falls abruptly as the audience departs, generating a "shark-fin" profile for CO_2.

Figure 5.1b shows CO_2 measurements and audience numbers for a day on which four films were screened, "Hunger Games 2", "Dinosaurs 3D" and "Buddy" twice. Those films with higher attendance have correspondingly higher CO_2. Figure 5.1c displays the CO_2 profile of a single film, "Hunger Games 2". Clearly the CO_2 trace does not increase smoothly with time, as would be expected from a constant emission source, but rather small peaks are discernable despite the cinema ventilation rate remaining constant. These CO_2 peaks would be generated if the audience's pulse and breathing rate were momentarily increased in response to scenes in the film. Figure 5.2 shows measurements from four showings of "Hunger games 2" on sequential days between December 26th - 29th with attendances of 87, 96, 104, and 186 people respectively. Two distinct peaks in CO_2 occurring around 15:00, highlighted by the red vertical lines, are visible on all days, indicating that the physiological response induced in each audience is reproducible. The pattern of CO_2 peaks shown in Figure 5.1c was characteristic of the film "Hunger Games 2" and in many cases it was possible to identify the different films from the CO_2 profile by eye.

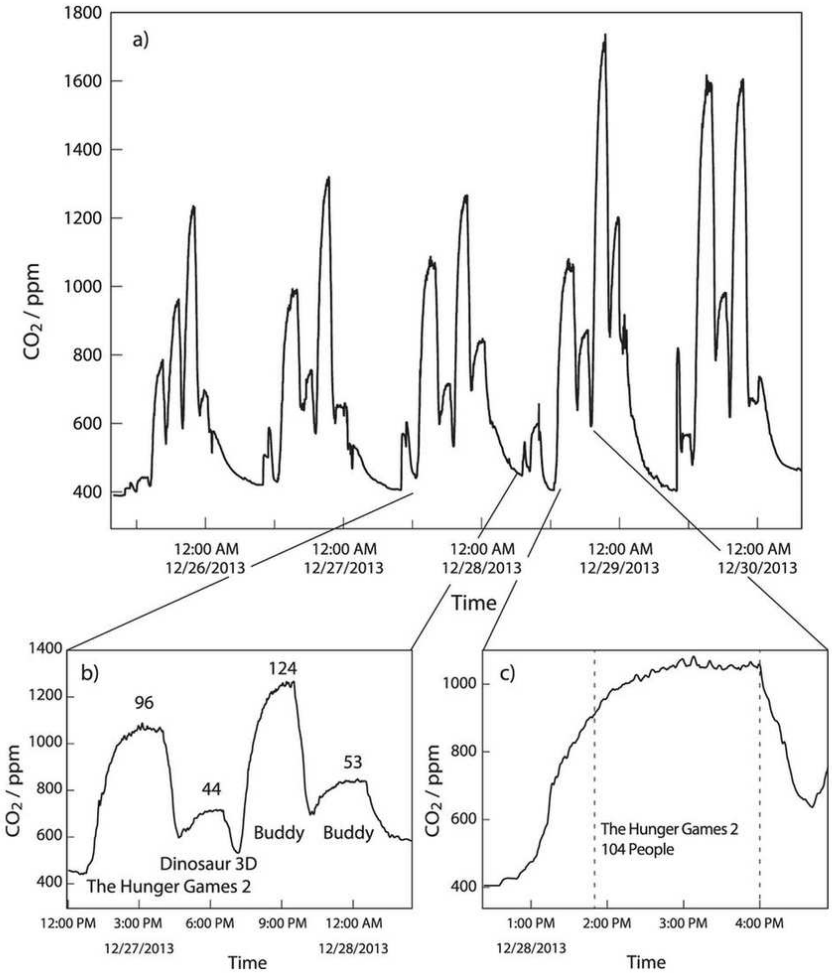


Figure 5.1.: Selected sections of the CO_2 measurements, (a) 5 days, (b) 1 day and (c) 1 film. The numbers above the peaks indicate the number of people in the audience.

The mixing ratios of isoprene (C_5H_8) and acetone (C_3H_6O), which are among the most abundant exhaled organic trace gases,[24, 75] are shown with CO_2 for four film screenings in Figure 5.2. Acetone is a soluble gas (in blood and water) that has been linked to fat catabolism, while isoprene is an insoluble gas linked to cholesterol synthesis.[75, 131] In Figure 5.2 peaks can be seen in the isoprene trace and to a lesser extent for acetone, although acetone mixing ratios were twice as high. Isoprene levels in the cinema are similar to levels reported from aircraft flying low over the pristine Amazon rainforest (1-3 ppb)[26] while acetone levels generated by the audience (~8 ppb) are approximately twice that found in forested environments[164] and city air.[50] The two distinct peaks around 15:00 previously noted in CO_2 are also visible in isoprene, and additionally a further large isoprene peak is observed at the end of each film (16:00). Breath analyses of individuals on an ergometer have shown that isoprene can be stored in muscle tissue, and that limb movement increases isoprene in breath.[76] The mass exodus of people at the

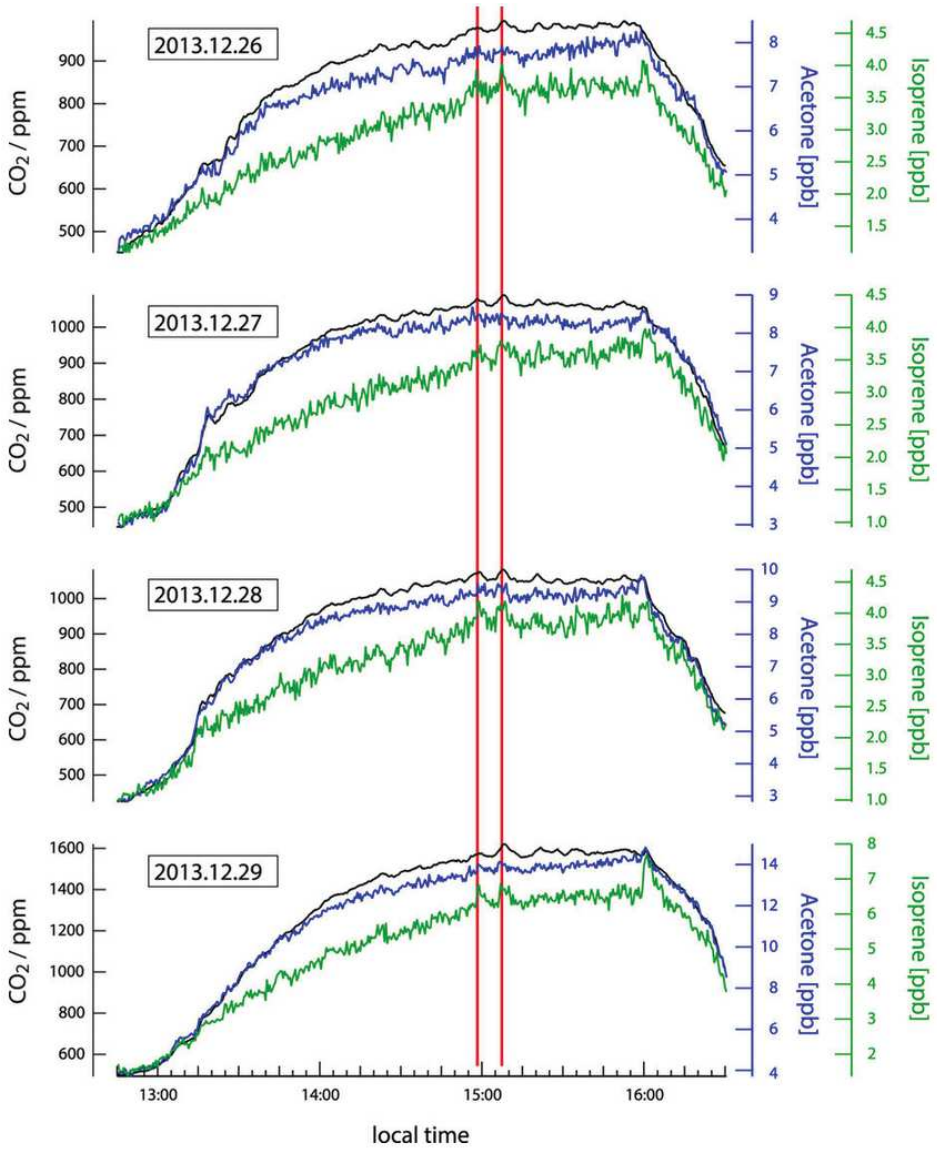


Figure 5.2.: Measurements of CO_2, isoprene and acetone taken during four separate screenings of "Hunger Games 2".

end of the film is therefore the likely cause of the isoprene peak at 16:00 coincident with rapidly falling CO_2. However, the two other outstanding peaks in isoprene appear during the film when the audience is seated (15:00 and 15:10). These times correspond to key moments in the film when the heroine's dress catches fire and when the final battle begins. Previous studies have indicated that breath holding[133] and twitching muscles[76] could potentially enhance isoprene emission over acetone. Another possibility is that isoprene is linked to cortisol production via cholesterol. Whatever the mechanism behind the release, the peaks in isoprene were reproduced in all four screenings of the film at the same time, meaning that each set of cinemagoers broadcasted chemicals into the air in synchrony to on-screen events.

Table 5.2.: Content labels

Label	Sub-label	Relative Frequency
Everyday Life		0.025
Dream		0.016
Landscape		0.04
Conversation		0.68
	Aggressive	0.008
	Conv. Main Actor	0.321
Action		0.141
Death		0.022
Running		0.031
Recovery		0.001
Laughter		0.004
Sleeping		0.001
Blood (violence)		0.028
Sex		0.003
Kissing		0.009
Crying		0.007
	Main Char. Cry	0.006
Injury		0.008
Sudden shock		0.026

Table 5.3.: Genre labels

Label	Sub-label	Relative Frequency
Suspense		0.283
	Chase	0.002
	Hidden Threat	0.005
	Hiding	0.002
Comedy		0.054
Romantic Comedy		0.002
Mystery		0.002
Romance		0.014
Drama		0.019

To determine whether causal links exist between levels of all chemicals measured and events in the film, it was necessary to annotate the films with scene content labels. A set of scene labels (Table 5.2 and 5.3) was defined based on genres in the IMDb database (e.g. comedy), on objective subheadings (e.g. chase) and psychological studies (happy to sad and excited to calm). These labels were applied to the films by ten individuals independently (see method for details). All data were then statistically normalized and random forests were constructed for each mass and CO_2, for each 30 second timestep within a 10 minute window, and for each label.[151] Each random forest based model was generated based on a randomly selected subset of two thirds of the data and then evaluated on the remaining third. This procedure was then repeated 15 times, using the Mainz Mogon supercomputer. A set of models were trained in a process called backward prediction to determine how well the present label was predicted by the future mass (in the next 5 minute time window). Figure 5.3a shows film scene labels plotted against AUC (Area Under Curve, see method) which expresses the ratio between true positives (when the model correctly predicted labels based mass decision trees) and false positives. A random prediction produces an AUC value of 0.5. Many of the labels showed a significant relationship with measured masses. The highest AUCs observed were for the labels “injury” (0.85), “hidden” (0.83), “mystery” (0.81) and “hiding” (0.79), all of which were subcategories of the label “suspense” which itself showed an AUC value of 0.75. The label comedy was also predictable based on the measured chemicals (AUC = 0.78). In contrast, the label “chase” (AUC = 0.55) could not be predicted by the model.

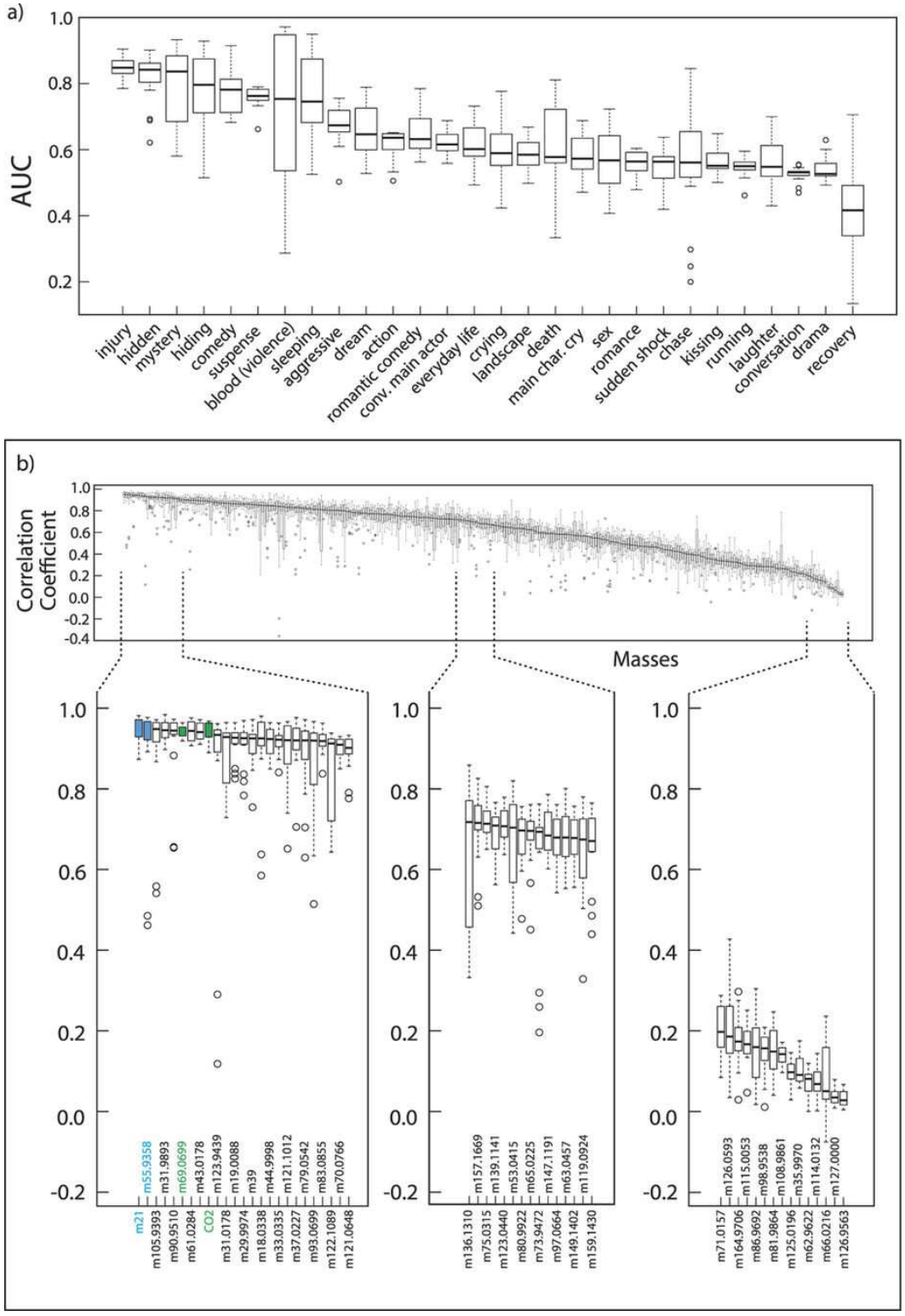


Figure 5.3.: Shown are the results when two thirds of the whole film screening dataset is randomly selected (15 times) and the resultant model tested on the remaining third.

The boxes indicate the extent of 25 percent of the data either side of the median (solid line). The dashed vertical line represents the lowest/highest datapoints that are still in the 1.5 interquartile range while the circles are outliers. Figure 5.3a shows AUC which expresses the ratio between true positives (when the model correctly predicted labels based on mass decision trees) and false positives (backward prediction). A random prediction produces an AUC value of 0.5. Figure 5.3b shows the ability of an individual mass to be predicted by the labels (forward prediction). The performance of this prediction versus the real value for VOC mixing ratios is given as the Pearson's correlation coefficient (r). High correlation coefficients indicate the predictive model was successful for that particular species, and not that all species with high correlation coefficients are inter-correlated.
In parallel we investigated the ability of an individual mass to be predicted by the labels (forward prediction). The performance of this prediction versus the measured mixing ratio is given as the Pearson's correlation coefficient (r) in Figure 5.3b. Strong correlation was found between model predicted and measured CO_2, as well as for the predicted and measured water sensitive reagent clusters m21 and m39. Both water and CO_2 are introduced to the cinema primarily by breath. Among the best correlated masses was isoprene ($r = 0.91$), which is presented qualitatively for the film "Hunger Games 2" in Figure 5.2. Some masses with high correlations have not been observed or identified in previous studies (e.g. 105.93, $r = 0.92$) while other masses exhibit no significant correlation.

Table 5.4 shows the best correlated masses and labels based on backward prediction. A filter of AUC > 0.5 and significance level <0.05 was applied to all data. "Significance" (Sig.) here is the result of a statistical T-test (between an evaluation based on all masses and an evaluation with one mass omitted, this mass is given in Table 5.4). Therefore higher AUC and lower significance values indicate stronger potentially causal links. The labels with the highest overall causal link to the measured species were "injury" and "comedy". Among the chemicals linked to injury scenes are methanol (m33.0335), acetaldehyde (m45.0335), 2-furanone (m85.0284), and butadiene (m55.0580). These compounds have all been previously detected in human breath.[24] Although the masses m100.9380 and m73.9472 were also significantly linked, no plausible identification could be made based on combinations of C, H, and O. Curiously, the mass m374.08 also shows a causal link to injury scenes despite being associated with polysiloxane which is found in cosmetics and conditioning shampoo. This may be related to emotionally induced body temperature variations rather than to breath. The film labels "chase" and "romance" both did not show significant causal links with any measured masses.

5.3. Discussion

Interestingly, the two film scene labels with the most significant linkage to chemicals measured were "suspense" and "comedy". These could be interpreted as an evolutionarily advantageous alert/stand-down signal, if perceivable by others.[134] Humans possess a very well developed sense of smell,[125] and new evidence suggests that recall is more

Table 5.4.: Film labels and masses with significant causal links are shown (Injury, Comedy, and Mystery) and two examples where masses and labels were not linked (Romance and Chase). The AUC value for "none" means the result with the complete dataset, and the values below are AUCs when the stated mass is removed from the model. Significance and AUC are given for each mass as well as an elemental formula and possible molecular identities based on previous measurements from human emissions summarized by de Lacy Costello et al.[24] The abbreviations refer to where the species were previously measured Br=Breath, Sk=Skin, U=Urine, F=Faeces, Bl=Blood and M=Mucus).

Injury				
Mass	Sig.	AUC	Formula	Possible ID/Comment
none		0.84929		
m374.082	0.004	0.81808		siloxanes
m73.947	0.015	0.8284		
m85.028	0.018	0.82524	$C_4H_4O_2$	2 (5H)-furanone (Br)
m105.034	0.03	0.81353	C_4C_8OS	3-(methylthio)-propanal (F, U, M)
m100.938	0.031	0.81562		
m40.974	0.035	0.82323		
m45.034	0.04	0.82347		Acetaldehyde (F, U, Br, Sk, M, Bl, Sa), Ethylene oxide (F, Br)
m55.058	0.04	0.83322		Butadiene (Br), Butyne (Br)
m33.034	0.044	0.82877		Methanol (F, Br, M, Bl)
Comedy				
none		0.77843		
m235.208	0.01	0.75878	$C_{15}H_{26}N_2$	
m111.080	0.031	0.7636	$C_7H_{10}O$	1,3-cyclohexadien-1-yl methyl ether (Br), 2-ethyl-5-methylfuran (F, U, Br), (E, Z)-2,4-heptadienal (M), 3-methyl-2-cyclohexen-1-one (U, Br), propylfuran (Br), 2,3,5-trimethylfuran (U, Br)
m121.065	0.045	0.76266	C_8H_8O	Acetophenone (F, U, Br, Sk, M, Sa), 2,3-dihydro-1-benzofuran (Br), 4-methylbenzaldehyde (F), phenyl acetaldehyde/phenylethanal/benzene acetaldehyde (F, M)
Mystery				
none		0.79193		
m217.204	0.024	0.7327	$C_{15}H_{20}O$	a-hexyl cinnamaldehyde (Sk)
m108.959	0.03	0.69253		
m159.143	0.04	0.69335	$C_9H_{18}O_2$	1-methylhexyl acetate (Sk), isoamyl butanoate (Br), heptanoic acid, ethyl ester (F), hexanoic acid, propyl ester (F), 3-methylbutanoic acid, butyl ester (F), 2-methyloctanoic acid (Sk), 2-methylbutyl 2-methylpropanoate (Br), nonanoic acid (U, Br, Sk, M, Sa), pentanoic acid, butyl ester (F), propanoic acid, hexyl ester (F)
Romance				
none		0.55738		
m95.049	0.157	0.54349	C_6H_6O	
m79.002	0.165	0.54388		
m70.077	0.289	0.54591	$^{13}CC_4H_8$	Isotope of isoprene
Chase				
none		0.55248		
m122.109	0.128	0.47477	$C_8H_{11}N$	
m100.084	0.135	0.47568	C_5H_9NO	
m164.971	0.169	0.47155		
m135.030	0.175	0.5066	$C_8H_6O_2$	

effective,[82] and our perception of faces changes with odours present.[146] Therefore the chemical accompaniment generated by the audience has the potential to alter the viewer's perception of a film.
There are several important consequences of our finding that human beings respond to audiovisual cues through breath emissions. Firstly, in the field of medicinal breath analysis, where chemical markers for diseases such as cancer are being sought,[75] emotionally induced emissions have the potential to confound disease marker identification. The strong response found here for "suspense" suggests that a patient's state of anxiety should be taken into account in future medicinal breath studies. These findings also have obvious industrial applications where an objective assessment of audiovisual material is sought from groups of people, for example, in advertising, video game design or in film making.

5.4. Method

5.4.1. Cinema/Movie Theater

All data were recorded at the Cinestar Cinema complex in Mainz (Figure 5.4a), Germany between 1st December 2013 and 14th January 2014. Of the 14 screen multiplex, two separate screen rooms were used (see Figure 5.4b, Cinema 2 capacity 230, and Cinema 7 capacity 230). During a film the entrance doors were closed and ambient air was circulated from outside into the room through vents under the banked seating and out via ceiling mounted openings so that the screening room was flushed entirely circa 6 times per hour. The measurement instruments (PTR-TOF-MS and the CO_2 detector, see below for details) were located outside the screening room (to avoid possible noise disturbance), in a technical room that contained the outgoing air vents (75 × 75 cm square stainless steel) and associated control systems for all auditoriums, see Figure 5.4c. An inlet was inserted into the midpoint of the exit flow vent and a 10 L/min flow was drawn through $\frac{1}{4}$" OD (0.625 cm) Teflon line continuously, see Figure 5.4d. The films viewed and the number of screenings are given in Table 5.1. This is a study of ambient air and the chemical changes within it caused by entirely anonymous groups of people in a public space. No personal data concerning the cinemagoers was collected, no individuals identified, only the number of people present were recorded by way of the ticket sales.

5.4.2. Proton transfer reaction time-of-flight mass spectrometer

Volatile organic compounds (VOCs) were measured using a commercial PTR-TOF-MS (proton transfer reaction time-of-flight mass spectrometer, PTR-TOF-MS 8000, Ionicon Analytik GmbH, Innsbruck, Austria).[48, 142] The measurement technique is based on the low pressure (ca. 2 mbar) protonation of molecules with a proton affinity higher than

Figure 5.4.: (a) The Cinestar Cinema in Mainz, Germany, (b) The 230 seat capacity cinema audioreum, (c) the air ventilation system, (d) insertion of the Teflon inlet into the 75 × 75 cm ventilation system. (a,b) are reproduced with permission from Cinestar.

water by H_3O^+ ions (691 kJ mol^{-1}) that are generated in a hollow cathode discharge chamber flushed with water vapour. All protonated molecular ions are accelerated by an electrical field to the same kinetic energy such that the resultant velocity of the ions depends on the mass-to-charge ratio. Hence, the time-of-flight is used to measure the velocity, from which the mass-to-charge ratio can be determined. The TOF was configured in the standard V-mode with a mass resolution of approximately 3700 m/z. Mass spectra were collected ranging from m/z 10–400 with a TOF acquisition sampling time per channel of 0.1 ns. The instrument was operated with a drift pressure of 2.20 hPa (E/N 137 Td) and a drift voltage of 600 V. For mass calibration, 1,3,5-trichlorobenzene was used as an internal standard by permeating 1,3,5-trichlorobenzene into a 1 mm section of 1/8" (1.58 mm) Teflon tubing used in the inlet system. Data post-processing and analysis was performed by using the program "PTR-TOF DATA ANALYZER", which is described elsewhere.[103] The PTR-TOF-MS was calibrated with a commercial pressurized gas standard mixture (Apel-Riemer Environmental Inc., Broomfield, USA) of known mixing ratio. The overall uncertainty was 15 percent. The calculated detection limit (3σ of the noise) of identified masses was between 15 ppt and 155 ppt. Signals were normalized to H_3O^+ ions and the first water cluster $H_3O(H_2O)^+$ by means of the following formula:

$$[\mathrm{R+}]_{\mathrm{ncps}} = [\mathrm{R+}]_{\mathrm{cps}} \cdot P \cdot 10^6 \cdot 296.15/(([\mathrm{m21}] \cdot 500 + [\mathrm{m39}] \cdot 250) \cdot T \cdot 2) \qquad (5.1)$$

here $[R^+]$ ncps is the normalized counts per second, $[R^+]$ is the reagent ion, P the pressure, T the temperature, [m21] the counts per second of the ^{18}O isotope of H_3O^+ and [m39] the counts per second of the ^{18}O isotope of the first water cluster of the primary

ion. The signal is normalized to a temperature of 298.15 K and a pressure of 2 mbar. The humidity dependence of the PTR-TOF-MS sensitivity was tested for a suite of compounds including key breath species such as isoprene and acetone shown in Figure 5.2. The sensitivity was weak, varying in the order of 3% for the ambient conditions in the cinema and therefore we can exclude humidity dependent variations in sensitivity as the cause of the peaks shown.

5.4.3. Carbon Dioxide (CO_2) measurement

CO_2 was measured at 1 Hz using a commercially available Li-COR Li-7000 system. The Li-7000 monitor was calibrated using a standard containing 509 ± 10 ppmv of CO_2 ppmv (Air Liquide, Germany) before and during the campaign. The instrument specifications state that the response is linear up to 3000 ppmv. Post campaign the linearity of the response was confirmed to 3400 ppmv using a second standard gas (10 percent CO_2, Air Liquide, Germany).

5.4.4. Film scene annotation

In order to assess the data for relationships between film scene content and trace gas behavior it was necessary to annotate the film scene content at high time resolution, from a set of preselected labels. Although several approaches to film scene annotation have been reported, including scene change frequency and both audio and visual cues,[47, 103, 169, 171] as yet no standardized procedure exists. Suitable independently derived time resolved annotations were also not available from film censor boards nor from the subsequently published film DVDs. Instead, ten volunteers individually viewed the films and allocated descriptor annotations as a function of the film duration using a custom made interface. Each film was labelled at least five separate times. Three different types of scene labels were used. The first set was general in nature and described the film genre using terms from the Internet Movie Database (IMDb). These included terms such as “comedy”, “suspense” or “romantic.” The second set was more specific and referred directly to the scene content such as “chase”, “laughter” or “kiss”, “house pet” or “injury”. These terms were kept deliberately objective to minimize potential labelling differences between individuals caused by personal perception. Finally, we have adopted an emotional assessment scheme that has been previously used by psychologists.[15] It consists of two separate five point scales, one ranging from happy to sad and the other from excited to calm. The labels produced by the individual volunteers were then averaged and used only when two thirds of the individuals agreed. The labels were created to match the datapoint frequency (1 every 30 seconds). A full list of scene labels is in Table 5.2,5.3 and a comprehensive description of all data mining approaches applied to the dataset given by Wicker et al.[151]

5.4.5. Data Mining

This study was designed to determine whether causal links exist between levels of volatile organic compounds and CO_2 emitted in a cinema auditorium and events in the film. While it is easy to examine the variance with time of a single molecular species for a single film by simple graphical methods (see for example Figs 1 and 2), to analyze the entire suite of measured masses (including unidentified mass species) at thirty second intervals with all the labels from all the films for causal relationships and possible interdependencies requires a more sophisticated and systematic data mining approach. Full details of the data mining algorithms applied are given by Wicker et al.[151], however, the generalized approach is summarized below. Data mining algorithms were applied to analyze the VOC and label data within a 10 minute window around a given measurement datapoint (5 minutes backwards and 5 minutes forwards). The first method applied was forward prediction, whereby the VOC mixing ratios are predicted based on regression from past VOC mixing ratios and the film labels. The second method was termed backward prediction, as it used VOC changes ahead of a given point in time to predict the current associated label. In order to evaluate the coherence of the two types of models, the forward prediction model and the backward prediction model, we used the predictions of the forward prediction model as an input to the backward prediction model and compared the resulting predicted values with the actual values. The overall product of the backward prediction are tables of VOC signal intensities (measured as mass-to-charge ratios in the mass spectrometer) that are associated with a given label and the error in the prediction expressed as the area under the receiver operating characteristic (ROC) curve (AUC, sometimes also called AUROC, see Table 5.4) and a significance. The AUC expresses how well a classifier (in this case the label) ranks the cases of one class before those of another class (in our case: those of one scene label before those of all others). An AUC value of 1 would mean that the label was predicted perfectly from mass signals, while a value of 0.5 indicates that the predictive performance was equivalent to a random selection.[14] The p-value results from a statistical test that compares the performance of a machine learning model using all masses as input to the performance of a model using all but one mass as input. The difference between these two cases is tested using a corrected paired t-test.[104] The t-test returns a significance measure in terms of p-values, the lower the p-value, the more probable is a relationship between the left out mass and the target label. Whereas in most cases, an adjustment like Holm-Bonferoni should be performed on the tests, this is not necessary in this case, as we only searched for indications for further analysis, which we also can get from uncorrected values. The results of the two (significance level and AUC in Table 5.4) expresses the significance of the relationship with low number of p-values and high numbers of AUCs indicating higher degrees of dependence.

5.5. Acknowledgements

We are very grateful to the Cinestar company for permission to use their facilities. In particular we thank Jochen Wulf (manager of the Mainz Cinestar) and his entire team for their enthusiastic support for this project. Finally, JW acknowledges the UK's flagship film podcast for inspiration in conceiving the overall project (L-O-2-J-SON-I-Saacs).

5.6. Contributions

J.W. conceived the experiment, made measurements and wrote the paper. B.D., C.S., E.B. and T.K. made the measurements, calibrated the systems and worked up the data to ambient concentrations. S.K., J.W. and N.K. conceived and performed the data mining analysis.

6. Can the age classification of films be made based on audience breath-chemical emissions?

C. Stönner[1], A. Edtbauer[1], B. Derstroff[1], E. Bourtsoukidis[1], T. Klüpfel[1], J. Wicker[2], J. Williams[1]

[1] Max Planck Institute for Chemistry, Mainz, Germany
[2] Department of Computer Science, The University of Auckland, Auckland, New Zealand

Manuscript submitted to PLOS ONE

Abstract Humans emit numerous volatile organic compounds (VOCs) through breath and skin. The nature and rate of these emissions are affected by various factors including emotional state. Previous measurements of VOCs and CO_2 in a cinema have shown that certain chemicals are reproducibly emitted by audiences reacting to events in a particular film. Using data from films with various age classifications, we have studied the relationship between the emission of multiple VOCs and CO_2 and the age classifier (0, 6, 12, and 16) with a view to developing a new chemically based and objective film classification method. We apply a random forest model built with time independent features extracted from the time series of every measured compound, and test predictive capability on subsets of all data. It was found that most compounds were not able to predict all age classifiers reliably, likely reflecting the fact that current classification is based on perceived sensibilities to many factors (e.g. incidences of violence, sex, antisocial behaviour, drug use, and bad language) rather than the visceral biological responses expressed in the data. However, promising results were found for isoprene which reliably predicted 0, 6 and 12 age classifiers for a variety of film genres and audience age groups. Therefore, isoprene emission per person might in future be a valuable aid to national classification boards, or even offer an alternative, objective, metric for rating films based on the reactions of large groups of people.

6.1. Introduction

With box office revenues worldwide estimated to be around 40 billion US dollars, the global film industry is an important element of many national economies. Once a film is recorded and edited it is must be classified prior to distribution to the cinemas. Movie

classification serves to protect children from unsuitable media content and to inform consumers, particularly parents, of the film's subject material. This classification is made at the national level by an independent regulator according to guidelines based on the legal framework of the individual country. The regulator assigns a rating to the film that reflects the public's sensibility to the film's content, ranging from unrestricted (suitable to all) up to adults only (typically 18 years old). The division of the classification system into age groups varies greatly from country to country. For example, Germany uses 0, 6, 12, 16, 18, while the United States has G (general audiences), PG (parental guidance suggested), PG-13 (parents strongly cautioned), R (restricted) and NC-17 (no one 17 and under admitted). India the world's most prolific film maker, uses U (0 to 11), UA (to 17) and A for adult. The classification process is complicated by the numerous influencing factors that must be considered together before the age classifier can be assigned, such as the degree of violence, sex, antisocial behaviour and bad language. Furthermore, public opinion on certain aspects of the classification guidelines may change with time requiring the regulator to revise their guidelines regularly. Ultimately, the classifying authority expresses a subjective assessment on behalf of the public in the form of an age limit. On some occasions this can be a contentious decision as a film maker seeking a larger market for their film may consider their work suitable for a broader audience than the classifying agency.

Clearly, it would be helpful to classification authorities if objective data based methods could be used to support the decision. Recently it was shown that cinema audiences emit chemical signals into the surrounding air in response to specific scenes in a film. Moreover, the sequence of signals over time was reproducible over multiple screenings of the same film. The effect can be most easily understood in terms of carbon dioxide (CO_2), which makes up circa 4% of exhaled human breath. Cinemas are ventilated continuously with outside air containing circa 0.0004% CO_2 so that when an audience is present the CO_2 level rises smoothly until an equilibrium is reached. However, when audience pulse and breathing rates increase momentarily in unison, in response to a particularly exciting scene, a peak in the CO_2 is generated which can be detected in air vented from the cinema. Current air measurement technology allows, in addition to CO_2, several hundred volatile organic compounds to be measured at high frequency (every 30 seconds). In the aforementioned study, it was found that certain chemicals corresponded to specific scene types, with suspense and comedy scenes being best characterized. The chemical response measured in the "crowd breath" represents the reaction of a large group of people to the scenes shown. This information could be potentially very useful in film classification as the chemical information is a direct, non-invasive measure of how a large group of people react to particular scenes and to the film as a whole. It is easy to imagine that the variability in the CO_2 trace, the number of peaks in the individual VOCs or the absolute amounts of the chemicals emitted per person are all possible indicators of the group response.

Recently, computerized systems evolved to support the decision making of the age rating of a film by the committee. Most of these methods utilize the language (use of bad words) and the image properties (colour variance, shot length) to classify the films[18, 67] but do not take into account the human reaction to the film. In this study we systematically

examine the feasibility of using CO_2 and over 60 VOCs measured in air ventilating from a cinema to classify films. The assessment is based on 135 screenings of 11 different films collected over 8 weeks from two separate cinemas involving more than 13000 people. Our approach involves a random forest model built with time independent features extracted from the time series of every measured compound for every film. These features include for example peak height, peak width and the number of peaks in a film normalized to its length.[92, 139] Finally, a permutation test was performed to test the resulting performance measures (area under ROC curve) of the original model versus the ones calculated from randomized class labels.[108]

6.2. Materials and Methods

6.2.1. Cinema measurement

We are very grateful to the Cinestar company for permission to use their facilities. No specific permission was required. Individual audience members were neither harmed or identifiable in the gas mixture and therefore the measurements were not subject to ethical approval.
The measurements were conducted in the multiplex cinema Cinestar in Mainz, Germany (located at 49° 59' 37.511" N 8° 16' 45.548" E) in two different screening rooms for approximately four weeks during the winter 2013/2014 and winter 2015/2016. Over the 8 weeks of measurement, 11 different films were shown multiple times resulting in a total of 135 separate screenings. Table 6.1 summarizes the films measured, categorized according to the German film classification system age recommendations "FSK" ("Freiwillige Selbstkontrolle der Filmwirtschaft" meaning voluntary self-regulation) along with the number of screenings. The average number of people present at each screening is given in the supplementary Table B.1. It can be seen that each age recommendation class was attended by approximately the same amount of people.
For this study, the German motion picture rating system was used dividing the films in 5 categories. Unrestricted films are classified as "FSK 0", films released to 6-years-old and over as "FSK 6", films released to 12-years-old and over "FSK 12", films released to 16-years-old and over "FSK 16" and films allowed only to adults "FSK 18". During the period of measurement, no film with the age rating "FSK 18" was screened. Since children under 12 have a discounted ticket price, the proportion of viewers at a particular film under 12 could be taken from the ticket sales.
The two different screening rooms were approximately the same size with a seating capacity of 237 and 227 viewers respectively. The size of the screening rooms was 6500 m^3 and the rooms were continuously flushed with 1300 m^3h^{-1} fresh outside air. No internal influx of consumed air from the cinema was mixed with the fresh outside air. The entire exhaust air of the screening room was drawn through a 75×75 cm stainless steel ventilation shaft. The air from the exhaust shaft was measured in a separate technical room with a PTR-TOF-MS and a CO_2-Analyzer.

Table 6.1.: Summary of the measured films partitioned into the four different age recommendation classes.

Age recommendation	Movie	Number of screenings
FSK 0		
	Help I've shrunk my teacher	18
	I'm off then	33
Total		51
FSK 6		
	Buddy	10
	Walking with Dinosaurs 3D	12
	The Secret Life of Walter Mitty	13
Total		35
FSK 12		
	The Starving Games	2
	Hunger Games: Catching Fire	8
	Star Wars: The Force Awakens	34
Total		44
FSK 16		
	The Counselor	1
	Machete Kills	1
	Paranormal Activity: Ghost Dimension	3
Total		5

6.2.2. Proton transfer reaction time-of-flight mass spectrometer

The exhaust air of the cinema was measured with a PTR-TOF-MS 8000 (Ionicon Analytik GmbH, Innsbruck, Austria). The ionization of each analyte occurs via hydroxonium ions (H_3O^+) resulting in protonated positively charged ions. This transfer reaction proceeds only to molecules with a higher proton affinity than water (691 kJ/mol). Thus the system is blind to the main air components like nitrogen, oxygen and argon. The low energy involved in the protonation reaction results in small fragmentation of the analyte facilitating identification. A detailed description of the set up and the calibration can be found elsewhere.[132]

6.2.3. Data analysis

In total, 20% of the film screenings measured had to be discarded due to problems associated ventilation system checks around midnight (only in winter 2013/2014) and high VOC emissions from cleaning products in the morning masking some human emissions (only for pre-midday screenings).
The data analysis was divided up into a pre-processing step and a model building step. The latter includes the generation of instances and the partitioning into training and test sets. Finally, the resulting performance measures were compared to the results obtained from a permutation test.

Pre-Processing

The measured time series for the isoprene mixing ratio for one film ("I'm off then", "FSK 0") is shown as the black curve on the left side of Figure 6.1. As the audience enters the cinema, the mixing ratio of isoprene increases quickly at first and then steadily during the film before decreasing sharply at the end when the audience leaves the screening room. In the case of isoprene, the peak that can be seen at the end of each movie is caused by enhanced release of isoprene due to muscle contractions associated with standing up and walking out.[69, 77] This peak was discarded for the analysis by removing the last 5 minutes of each film in the data pre-processing step.
Zooming in, several peaks and valleys can be seen which re-occur at the same time in every screening of the same film.[155] The maximum mixing ratio of the VOCs measured for each film, positively correlates with the number of viewers attending the screening room. Therefore, the time series of the individual films were normalized to the number of viewers which is known from the ticket sales. The temporal behaviour of isoprene with an initial sharp increase, followed by a smooth steady enhancement and final rapid decrease was similarly observed for many other breath-borne compounds such as CO_2 and acetone. The increase in the mixing ratio (red curve in Figure 6.1) can be calculated using a box model assuming a constant emission rate during the film. Within the model the mixing ratio is dependent only on the inflowing and out-flowing air and the emission rate of the VOCs from the audience. A detailed description of the model can be found in the supplement. The modelled behaviour of the mixing ratio assuming a fixed emission rate (red curve in Figure 6.1) was subtracted from the measured mixing ratio (black curve on the left side in Figure 6.1). The resulting trace without the increasing trend was termed as the "residual time series" and can be seen on the right hand side in Figure 6.1.
The corrected time series was used to extract distinctive features comprising standard deviation, skewness and kurtosis of the time series as well as several features describing the occurrence of peaks in the time series. Additionally, the mean of the positive and negative values were included into the feature set to obtain an overall measure of change within the time series. The residual time series allows the comparison between the peak heights of different films. In one case all peaks were counted (single time step increase and decrease). In a second case, peaks were only counted exhibiting a sequence of a

]

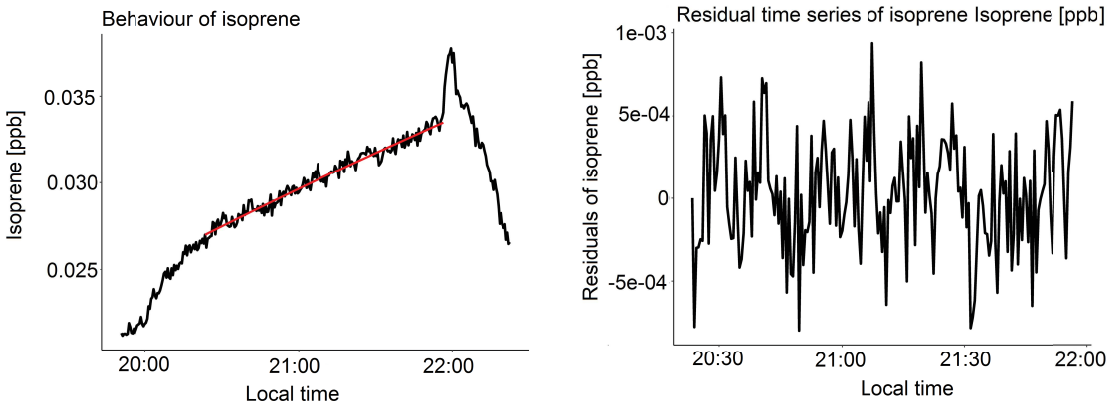


Figure 6.1.: Time series of isoprene. The left panel shows the mixing ratio of isoprene (in ppb) during the film "I'm off then". The black line shows the measured values and the red one the modelled mixing ratio assuming a constant emission rate of isoprene. The right panel shows the residuals obtained by subtracting the measured times series from the modelled one.

minimum of 3 consecutive increasing and decreasing steps. In the case of the peak height and peak width only the 5 highest and widest peaks with a minimum of 3 increasing and decreasing steps were taken and included in the feature set. In total 18 features were included in the feature set. The complete list of extracted features can be seen in Table 6.2. These features were extracted for each film and for each measured VOC. A separate model was built for each of the measured 66 VOCs.

Table 6.2.: Summary of the extracted features.

Extracted features
Moments: Standard deviation, Kurtosis, Skewness
Sum of all positive values
Sum of all negative values
Amount of peaks (all peaks and peaks with a minimum of 3 increasing and 3 decreasing steps) normalized to the film length
Occurrence of the first peak
5 highest peaks normalized per person
5 widest peaks normalized per person
sum of the 5 highest peaks
Sum of the 5 widest peaks

Model building

Instances were created for each molecule in the same manner using the four age ratings "FSK 0", FSK 6", "FSK 12" and "FSK 16". For the modelling process, the films were divided up into a training and a test set. For each age recommendation class, one film was chosen to be in the test set and the remaining films were put in the training set. In order to receive statistically meaningful results, the test set contained only films with 8 or more recorded screenings. An exception involves the age recommendation "FSK 16" because two films were measured only once ("The Counselor" and "Machete Kills"). These two films were always put into the same set (training or test set) and were evaluated together. Consequently, the other set includes "Paranormal Activity". This set up results in 24 combinations of different training and test sets (two possible films in "FSK 0", "FSK 12"and "FSK 16" and three possible films in "FSK 6").
For each training set a random forest model was constructed.[16] The random forest classifier was run with the default values for its parameter specifically the number of trees to grow was set to 500 and the number of variables randomly sampled at each split was set to 6 (number of variables divided by 3). This model was used to predict the age rating of the corresponding and unseen test set. The classifier performance was evaluated using Receiver Operating Characteristic (ROC)[37, 38] and Precision-Recall curves (PRC).[112] The ROC curve with its corresponding area under curve (AUC) value is frequently used in the machine learning community. However, this performance measure lacks interpretability when it comes to imbalanced data sets.[118] In our study the number of negative examples exceed the number of positive examples. For instance, a large number of false positives weakly enhances the false positive rate used in ROC. On the other hand, the precision value is affected by a larger amount because this value compares the false positives and the true positives. Between the ROC and PRC curves a one-to-one relationship exists, meaning that each point in one curve uniquely corresponds to one point in the other curve and vice versa.[27]

Permutation test

A permutation test was performed to check for spurious results. For this test a random age rating was assigned to each film in the training set and the initial class distribution was retained. The test set kept the original age ratings. From the resulting model, the area under curve from the ROC and the PRC were calculated. For each test set composition the training set labels where shuffled 50 times and the resulting 50 performance measures where compared to the corresponding original performance measures. Therefore, the cases in which the performance measure of the permutation test exceeds the one of the original test set composition were counted and divided by the total amount of permutations. The calculation of the p-value was according to Ojala et al.[108] Generally, a Holm-Bonferroni correction should be applied to the p-values to counteract the problem of multiple comparisons as in our case summarizing the models from all measured VOCs. In this study we did not apply this correction since we only searched for indications pointing to VOCs which might be useful for further analysis. Therefore, we

used the uncorrected p-values.

6.3. Results

The resulting AUC values are summarized in Table 6.3. The complete results with the corresponding standard deviation of the AUC values and p-values are given in the supplement (supplement Table B.2-B.4).
It can be seen that most of the VOCs show AUC values below or around 0.5 indicating a performance similar to a random classifier. The AUC value for CO_2 shows the highest value in the age class "FSK 12", whereas almost no significance is seen in the other categories. Isoprene shows AUC values above 0.7 for the age classes "FSK 0", "FSK 6" and "FSK 12". The AUC value for the age class "FSK 16" lies below 0.5. However, this class is hardly interpretable since we measured only 6 films for this class. This may pose a problem to the reliable prediction of this class. Therefore, in the following discussion the age recommendation class "FSK 16" was omitted.
In general, it can be seen that based on the AUC values several different compounds

Table 6.3.: Summary of several VOCs with the corresponding area under ROC curve for the different age classes. AUC values of 0.70 and higher or equal were highlighted in bold font.

Mass	Compound	FSK 0	FSK 6	FSK 12	FSK 16
CO_2		0.55	0.53	**0.75**	0.15
m31.0178	Formaldehyde	0.55	0.71	0.48	0.39
m33.0335	Methanol	0.50	0.62	0.36	0.26
m45.0335	Acetaldehyde	0.45	0.50	0.51	0.14
m59.0491	Acetone	0.55	0.63	0.56	0.13
m61.0284	Acetic acid	0.54	0.54	0.55	0.32
m63.0263	Methyl mercaptane / Dimethylsulfide	0.55	**0.76**	0.40	0.44
m65.0215		**0.74**	**0.79**	0.40	0.17
m65.0604		0.36	0.64	0.58	0.16
m67.0542		0.40	0.65	0.47	0.52
m69.0699	Isoprene	**0.84**	**0.74**	**0.70**	0.25
m83.0455		0.38	0.40	**0.70**	0.60
m95.0855		0.53	0.54	**0.70**	0.55
m137.1325	Sum of Monoterpenes	0.60	0.58	0.64	0.66
m235.2056		0.51	0.55	**0.73**	0.48
m355.0698	Fragment of Decamethylcyclopentasiloxane	0.54	**0.73**	0.59	0.38

are able to distinguish between one or more age classes. Isoprene, which is one of the main VOCs on breath, seems to be a potentially useful compound for the differentiation

of the age classes FSK 0, 6 and 12. Other compounds being able to predict only one age recommendation class are for example CO_2, formaldehyde and decamethylcyclopentasiloxane.

Figure 6.2 shows the average behaviour of the AUC values derived from the ROC

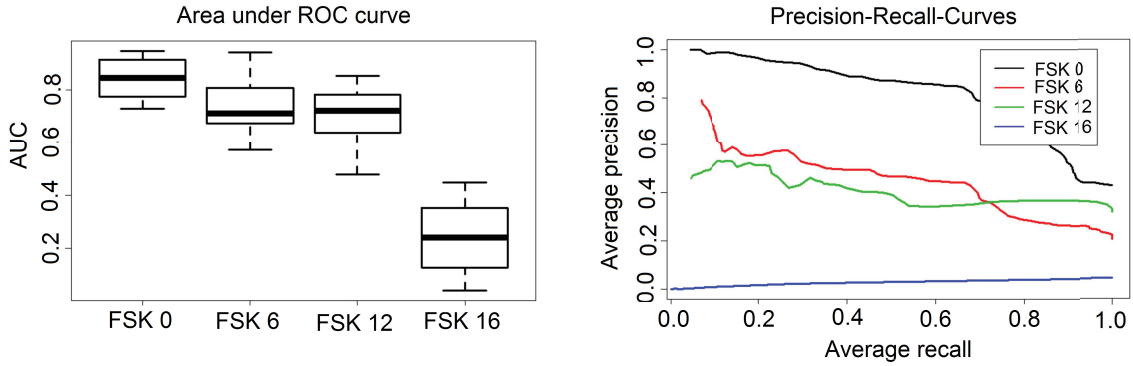


Figure 6.2.: Performance measures for isoprene models. The box plots of the calculated area under ROC curve values(left) and average Precision-Recall curves for all age recommendations(right). The performance measures were derived from the isoprene models. The box plots on the upper left panel in Figure 6.2 the thick black line in the middle of the box indicates the median value for each group. The box comprises the interquartile range (IQR) of the data and the whiskers define 1.5 times the IQR or the minimum and maximum if no points exceed the 1.5 time IQR.

curves and PRC for the age classes from 0 to 16 for isoprene. In general, 24 models were built for each measured VOC corresponding to the 24 different test and training set combinations. The plots in Figure 6.2 show the average value of the performance measures calculated from the 24 different test and training set combinations of isoprene. In case of the age class “FSK 0” all AUC values lie above 0.5 indicating a non-random classifier. For the age recommendations “FSK 6” and “FSK 12” some of the AUC values lie around the value of 0.5 (median AUC value ~0.70) indicating that some models trained and tested on particular sets cannot be predicted with a higher chance than a random classifier. The PRC shows the average curve over all models. The PRC curve shows a similar behaviour as the AUC values describing the age class “FSK 0” as the best predicted class followed by the age classes “FSK 6” and “FSK 12”. However, the age classes “FSK 6” and “FSK 12” exhibit higher average precision values of ~0.6 for lower recall values up to 0.2. The p-values from the permutation test for isoprene for the “FSK 0”, “FSK 6” and “FSK 12” are 0.01, 0.05 and 0.16 respectively. Consequently, for a significance level of 0.05 the null hypothesis for “FSK 6” and “FSK 12” cannot be rejected. Examining the variable importance for all 24 random forest models built for isoprene no specific feature was found to distinguish oneself from the others.

6.3.1. Different genre labels

In this section we examine the differences in classifier performance between films of the same age class but with different genre label. The genre labels for the films were taken from the International Movie Database (IMDb). For this purpose, the films of the age class “FSK 6” were selected due to their similar frequency in the number of screenings (10 films of “Buddy”, 12 films of “Walking with Dinosaurs 3D” and 13 films of “Walter Mitty”). Here the film “Walking with Dinosaurs 3D” was labelled as “action” whereas “Buddy” and “Walter Mitty” are “comedy” films.

Figure 6.3 shows the distribution of AUC values and PRC depending on the film in the

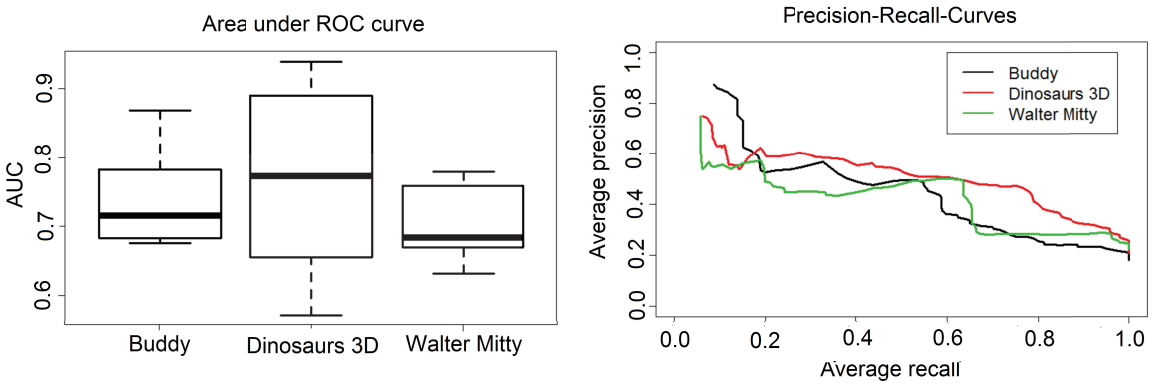


Figure 6.3.: Performance measures for isoprene models involving only the “FSK 6” films. Area under ROC curve and Precision-Recall-curves divided up into the different films in the age recommendation "FSK 6". The film “Walking with Dinosaurs 3D” was labelled as “action” whereas “Buddy” and ”Walter Mitty” are comedy films.

test set. On the right side the results are shown for the age recommendation “FSK 0” and on the left side the results for “FSK 6”. Note that one film was chosen in the test set putting the other two films in the training set. It can be seen on the left side that the highest mean AUC value ($\sim 0.77 \pm 0.13$) is obtained placing the film “Walking with Dinosaurs 3D” in the test set. This test and training set combinations also shows the highest standard deviation. Figure 6.3 compares the mean PRC of the three films in the age class “FSK 6”. In general, all three PRC curves seem to behave similarly despite using films with different genre labels. Evaluating the permutation tests separately for the three different films results in p-values for the film “Buddy” to be 0.06, “Walter Mitty” to be 0.04 and “Walking with Dinosaurs 3D” to be 0.11. In this case the p-values were calculated by comparing the original test set combinations in which the selected film appears with the corresponding randomized ones. These p-values are in accordance with the boxplot in Figure 6.3 showing the film “Walking with Dinosaurs 3D” with the highest standard variation. Thus, the chances are higher that the AUC values of the permutation test exceeds the values of the original one. It seems that some of the training and test set combinations do not predict the “Walking with Dinosaurs 3D” film

well and that the genre label has an effect on the prediction results. However, the film "Walking with Dinosaurs 3D" could be predicted comparably well keeping in mind that there is no other action film in the training set if "Walking with Dinosaurs 3D" is in the test set.

6.3.2. Different age of the audience

The ticket sales information provided the proportion of viewers younger than 12, to viewers 12 years or older. In the case of the age rating class "FSK 0" the films shown were aimed at quite varied audiences. The film "Help, I've shrunk my teacher" was classified as a "family" film by IMDb and the proportion of viewers younger than 12 was 64%. In contrast, the film "I'm off then" was attended only by viewers of 12 or older. The film "I'm off then" was more targeted to adults as it deals with a man on a pilgrimage in Spain. Remarkably, this age rating could be predicted with the highest AUC of 0.91 despite the difference in the average age of the viewers.

The difference in the AUC value between those two films with the averaged PRC

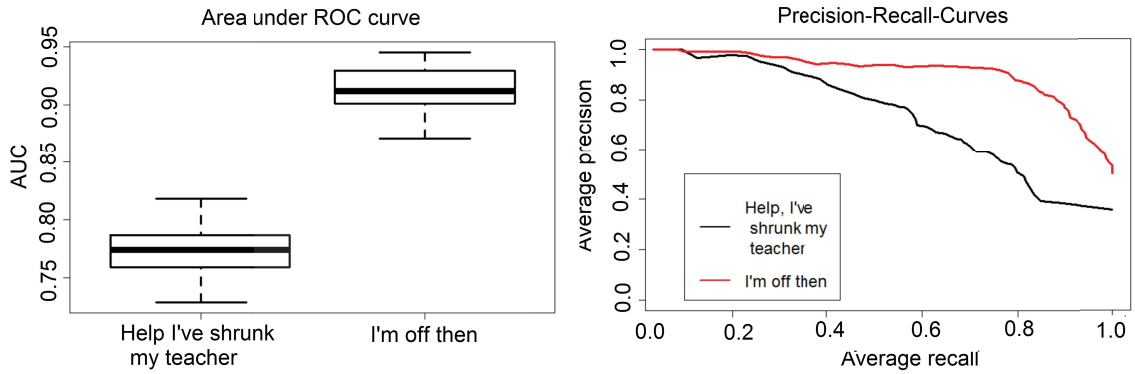


Figure 6.4.: Performance measures for isoprene models involving only the "FKS 0" films. Area under ROC curve and Precision-Recall-curves divided up into the different films in the age recommendation "FSK 0". The film "Help I've shrunk my teacher" was attended by a large proportion of viewers younger than 12 (64%) whereas the film "I'm off then" was only seen by viewers of 12 or older.

curve for each film can be seen in Figure 6.4 on the right side. Both PRC curves exhibit precision values higher than that of a random classifier. It can be seen that the classifiers perform worse, if the training set contains the film "I'm off then" and the test set "Help I've shrunk my teacher" than vice versa. Nevertheless, the age difference of the audience between those two films does not seem to worsen the classifier critically and the age class can be still predicted to a reasonable degree.

6.4. Discussion

In this study we have assessed whether the age classification of a film can be predicted based on variations of airborne chemicals measured in a cinema. Previous publications[151, 155] have reported correspondence between audio-visual stimuli and the emission of VOC from human beings. The aforementioned study reported that scenes labelled with "suspense" and "comedy" caused the audience to change their emissions of chemicals significantly. Intuitively then, we may think that these chemical changes may be related to the age classification given to the film. For example, films with horrific scenes will induce rapid pulse and breathing rates and hence higher and more variable levels of CO_2. Indeed, CO_2 values are effective in predicting FSK 12 films, probably because the films in this category were action films. However, our results show that most of the chemicals measured, including CO_2, do not reliably predict all the age classifications of the films (0, 6, 12). One reason for this may be that current age recommendations for films are not solely linked to the intensity of induced fear, or the audience's innate visceral responses to film content. Rather it is subjectively based on a synthesis of multiple aspects such as the degree and intensity of violence, sex, antisocial behaviour, drug taking and bad language. Provided that the audience's reaction to these aspects of the film are in some way reflected within the large chemical dataset, an alternative and objective age classification may still be possible. It is interesting to reflect that a chemical based approach as advocated here, would be based on directly measured responses from large test audiences, whereas the current scheme is based on a subjective appraisal of the film by relatively few people entrusted to reflect the general public sentiment.
Of all species tested we found that isoprene performed best in predicting the different age classifications. The highest AUC value was obtained for "FSK 0" (AUC value of 0.84 ± 0.07). In this age recommendation class, the two films had a different proportion of younger viewers (in "Help I've shrunk my teacher" 64% of the audience was younger than 12 and in "I'm off then" only viewers of the age of 12 or older attended). It is known, that the children emit significantly less isoprene than adults.[87, 132] Nevertheless, it was found that the age structure of the audience does not critically worsen the predictive power of the classifier and that the features displaying the structure of the films were able to distinguish this class from the rest. Lower precision values were reported for the age recommendation classes "FSK 6" and "FSK 12". For these two classes the p-value of the permutation test lies between 0.05 for "FSK 6" and 0.16 for "FSK 12". The age recommendation "FSK 6" included three films with similar frequency of two different genre labels (two comedy films and one action film). Again, this does not seem to influence the classifier's performance. In the case of "FSK 12", the average AUC values were 0.63 ± 0.10 if the film "Star Wars" is included in the test set (and the films "The Hunger Games" and The Starving Games" in the training set), and 0.77 ± 0.05 if the film "The Hunger Games" is included in the test set (the films "Star Wars" and The Starving Games" in the training set). The lower AUC (0.63) value and the higher corresponding p-value of the permutation test (0.25 versus the p-value of 0.08 including "The Hunger Games" in the test set) is likely due to the lower amount of training examples (12 screenings with "The Hunger Games" and the "Starving Games" in the

training set).
Isoprene is generated in the body during cholesterol-genesis[131] and stored in muscle tissue. Muscle movement causes stored isoprene to enter the bloodstream and then vent the body via the breath.[69, 77] It is interesting that that this species, rather than CO_2, can be used as a successful delineator for film classification, at least for FSK 0,6, and 12. Generally, it could be seen that isoprene was reproducibly emitted at higher rates at the same time point in the same film, even with different audiences. The height of peaks depicted in Figure 6.5 show smallest values for the age class "FSK 0". This could be because of the lower concentration of isoprene in the breath of children for the film "Help I've shrunk my teacher" or because of fewer suspense scenes in both films. Suspense scenes generally lead to increased heart and breathing rates as well as involuntary movement, all of which enhance the isoprene emission rate of the audience.

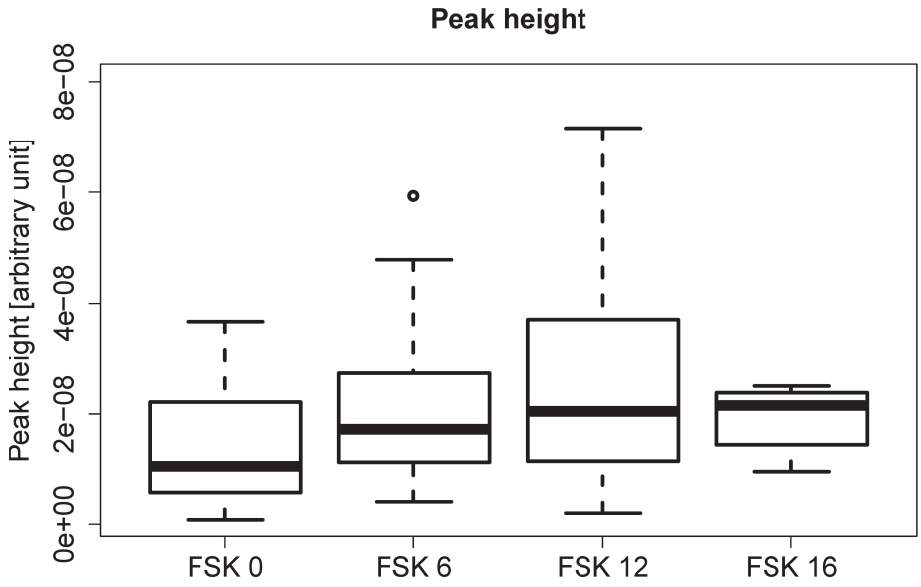


Figure 6.5.: Boxplot with the height of the highest peak for the different age recommendations.

We may speculate that in future isoprene may be used to objectively classify films, using the dataset shown here as the basis or conceivably, measurements from a test audience could be used by the classification board to aid in decision making in borderline cases.
The presented findings imply that features within the isoprene trace correspond to the pattern and the intensity of induced emotions in the film. This structure can be distinguished from the other classes even when the audiences between the films in the same class consists of different age proportions ("FSK 0") or the genre labels of the films are different ("FSK 6"). In general, the isoprene trace properties should reflect the subjec-

tive assessment of the rating agency (age classification). It would be interesting to see if isoprene acts as an indicator in other domains with multiple underlying stimuli like psychological stress. The perception of stress can also be caused by several environmental conditions and events.
In this study we used a random forest model built for each measured mass separately to predict the age recommendation of a film. Future work should involve combinations of different masses. As shown in Table 6.3 masses like m65.0215 or CO_2 show higher AUC values for the specific age classes like "FSK 6" (AUC of 0.79 for m65.0215) and "FSK 12" (AUC of 0.75 for CO_2) than isoprene. These masses might respond better to specific scenes or capture similar structures of films within the same age recommendation class. Thus, the combination of the features of these masses might help to reflect the structure of the film and to improve the classifier's performance. This approach requires a larger number of measured films because as in our case the classifier adopted the features of the few films in the training set very well (resulting in a perfect classification for the training set) so that the extrapolation to new films in the test set resulted in comparably low performance values. The addition of new features from other masses exacerbates this problem. Therefore, a new validation set should be used to choose the best combinations of these masses and to test them on an unseen set of films. Therefore, future datasets should include a larger number of films.

6.5. Conclusion

This study presents a framework to objectively gauge a film into an age classification system based on VOCs and CO_2 in cinema air. The evaluation of these compounds resulted in no single human generated volatile compound being able to distinguish all four age classes (0, 6, 12, 16). Problems arise because of the small number of available films within each age class, especially for the age class "FSK 16". Overall, the results of this first study are promising for isoprene. Perhaps in future metrics can be devised using combinations of isoprene and other VOCs to designate movie classification. This could be useful for the film industry which edits films to make them accessible for their desired target audience. It can be seen for the age recommendations "FSK 0" that the classification is based on the films in this class and not on the age structure (different target audience in "FSK 0"), so the classifying ability is not based on the lower isoprene emission from children. The concepts proposed here could be tested more thoroughly if more films are sampled. In particular, a larger suite of films with rating "FSK 16" and "FSK 18" would be interesting as they represent the extreme categories.

6.6. Acknowledgements

We are very grateful to the Cinestar company for permission to use their facilities. In particular, we thank Michael Djienes, Jochen Wulf, Constantin Maximilian Best, Axel Kessler, Richard Stotz, Stephan Lehmann and the entire support team for their enthusiastic support for this project.

7. Pattern identification in multivariate atmospheric time series

Abstract Atmospheric data time series typically contain multiple variables including meteorological information and the temporal behaviour of trace gases and particles. Such time series can contain categorical values for example wind direction data and air mass origin, along with continuous variables such as temperature, wind speed and the abundance of atmospheric species. The identification of patterns within this data set must take account these different parameter attributes and must be robust towards noise. In this study we have analysed data from two different ground based summer measurement campaigns that took place in Finland (HUMPPA-COPEC 2010) and Cyprus (CYPHEX 2015). A novel pattern identification method was applied to the data to extract periods of similar meteorological conditions within both campaigns. The pattern identification method has proven to be robust towards noise. Furthermore, such extracted data facilitate the interpretation of trace gas behaviour. Comparisons between different extracted patterns are presented. For further understanding the origin and fate of the trace gases the application of a regression tree model is demonstrated to qualitatively estimate the influence of meteorological variables on the measured organic compounds. These techniques can be applied generally to long term ground based datasets. It is shown that the pattern identification method is a powerful tool in connection with data mining methods for further understanding, particularly for interpreting VOC data where the species may have multiple sources with complex varying strengths.

7.1. Introduction

Temporal data mining can be divided into several tasks including the classification, forecasting, clustering and pattern identification.[43] This study focusses on pattern identification meaning the extraction of time periods of similar behaviour out of multivariate atmospheric time series. Statistical evaluation methods have been applied widely in research areas including atmospheric chemistry and mass spectrometric data.[13, 17, 36, 63, 94, 109, 161] However, the use of unsupervised pattern identification methods in atmospheric chemistry is sparse and is more widely applied in the field of economic research.

The approach applied in this study was first developed by Guimarães et al.[51, 52] and expanded by Mörchen et al.[100, 101] It provides a framework for identifying patterns in time series. The method is based on the discretization of the time series into labels such as “high”, “medium” and “low” and the subsequent search for sequences in this

discretized data. Promising results for the identification of patterns have been shown in the field of sports medicine. This work uses this approach for identifying patterns in atmospheric data.
Atmospheric data set can be quite complex. It usually contains time series of meteorological variables like temperature, sun radiance, relative humidity, wind speed and wind direction combined with time series of measured trace gases and particles. The time resolution of the measurements typically varies from steps of 1 minute to 1 or 2 hours depending on the measurement device. Here we evaluate meteorological data in conjunction with data from a proton transfer reaction time-of-flight mass spectrometer (PTR-TOF-MS).[12, 48] This device is able to measure hundreds of volatile organic compounds (VOCs) with a time resolution of a 1 minute. The time series of these trace gases can be quite noisy depending on the abundance of the trace gas, the integration time of the measurement and the associated sensitivity of the measurement device for this compound. The behaviour of VOCs in the field can be erratic. This mainly depends on the meteorological conditions such as highly varying wind directions and wind speeds which determines which sources impact the measurement.
The data used in this work was taken from two measurement campaigns in Finland in summer 2010 (HUMPPA-COPEC)[153] and in Cyprus in summer 2015 (CYPHEX).[30] Both campaigns took place for 4 weeks in summer and comprise two very different environments. In Finland the measurement site was located in a boreal forest at Latitude 61°51′0” N and Longitude 24°17′0” E (elevation 181 m above sea level) with its main focus on investigating the chemistry and physics of this boreal forest ecosystem. The air masses mostly originated from the south west (53.7%) with significant periods from the south east (20.7%) and north east (10.3%). In Cyprus the measurement took place on a 650 m high hilltop (34°57′0 N”/32°23′0” E) covered by sparse Mediterranean scrub vegetation. This measurement site was impacted primarily by northerly Etesian winds alternately shifting between air masses from over the sea originating from western Europe (for 33% of the time) and eastern Europe (for 67% of the time). The local wind direction was primarily south west (70%) with intermittent sea breeze effects.
The air masses encountered at the measurement sites contained multiple VOCs that originate from both biogenic and anthropogenic sources and some that are formed through photochemical processes. With increasing air mass age (local emissions versus long range transport) the primarily emitted compounds are gradually removed and oxidised to new species. By extracting periods of similar meteorological conditions allows the investigation of parameters such as temperature and relative humidity on the local emissions and the extent of photochemical processing. Secondly, pattern identification is an important tool to objectively classify time series data into time periods of similar behaviour and enables comparison between different meteorological regimes. Furthermore, we explore methods for evaluating the derived sequences and estimate the influence of the meteorological variables on the measured volatile organic compounds. Finally, the trace gas data includes many masses which have not been unequivocally identified. The behaviour and similarity to known VOCs for particular time periods can help identify those species.

7.2. Method

The data was obtained from two ground based summer field campaigns which took place in Finland (HUMPPA-COPEC) and the other in Cyprus (CYPHEX). For both campaigns data is available over a time period of approximately 3-4 weeks. The time resolution of the meteorological variables and the trace gases is 1 minute for the Cyprus data and 10 minutes for the Finland data. A detailed description of the measured compounds and deployed instruments can be found elsewhere.[30, 153]
The PTR-TOF-MS provides information about the molecular mass of a molecule. From this information the mass sum formula can be derived but the structural formula remains unknown. From previous field and laboratory studies several masses can be assigned to a particular molecule with confidence. These "known masses" were chosen in this study according their abundance and importance in atmospheric chemistry. Nonetheless it should be noted that other isomers or isobaric compounds can interfere with these mass signals, but here it is assumed that these known masses are the predominant contributors to the mass signal. In order to gain knowledge of the unknown masses these are compared to the known ones. It is assumed that a similar behaviour might indicated a similar source and fate in the atmosphere. Table 7.1 shows the masses which were known.

Table 7.1.: Known and calibrated masses for the PTR-TOF-MS used during the CYPHEX campaign and the PTR-MS used during the HUMPPA-COPEC campaign.

CYPHEX		HUMPPA-COPEC	
Mass	Compound	Mass	Compound
m33.0335	Methanol	m31	Formaldehyde + unknown compound
m42.0338	Acetonitrile	m33	Methanol
m45.0335	Acetaldehyde	m42	Acetonitrile
m59.0491	Acetone	m59	Acetone
m63.0263	Dimethylsulfide	m69	Isoprene
m69.0699	Isoprene	m71	Methyl vinyl ketone + Methacrolein
m71.0491	Methyl vinyl ketone + Methacrolein		
m73.0648	Methyl ethyl ketone		
m79.0542	Benzene		
m81.0699	Pinene fragment		
m93.0699	Toluene		
m107.0855	Xylenes		
m121.1012	Trimethylbenzene		
m137.1325	Monoterpenes		

For the pattern identification algorithm each continuous time series is discretized into a number of chosen labels. Smaller interruptions from another different label are filtered out, according to a predetermined time based threshold. Next the labels of all chosen time series are combined. In the final step, sequences are found in the previously combined and discretized time series. For the sequence mining, the data is divided into days. Thus days of similar meteorological conditions are assembled from pieces of the real time data. The following sections give a detailed description of the multiple data analysis steps.
The overall method is shown in Figure 7.1.

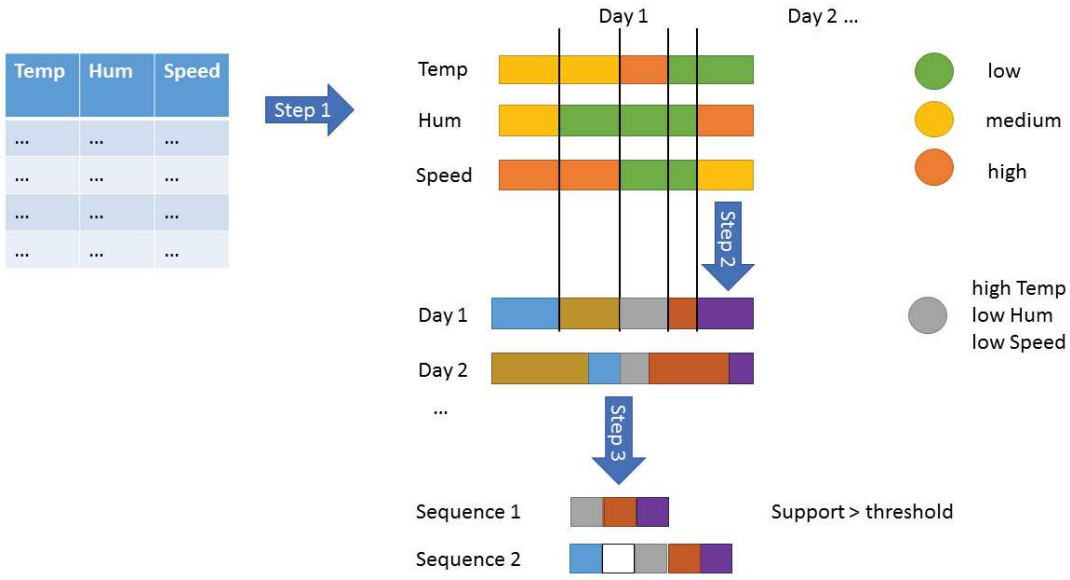


Figure 7.1.: Scheme of the applied method described in this section.

7.2.1. Discretizing the univariate time series

In this step the continuous univariate time series is discretized into several labelled segments. This discretization can be done using the whole data set at once or using a part of the data defined by a time window. The size of the time window can range between several hours to days. The reason for the application of a time window in this discretization step can be the elimination of a trend in the time series. More importantly, this window removes the information concerning the absolute value of the time series. Thus it is possible to search for the same patterns even if the data differs strongly in magnitude. Usually this time window is of the same length as the time window in which the sequences are mined in the last step. In this last step the length of an interval is chosen in which the sequence mining algorithm searches for frequent patterns. For example, the user wants to find sequences within of the maximum length of one day. Therefore, the window length is chosen to be 24 hours. Thus it makes sense to choose a

window length of the same size for discretization in order to remove some trend in the time series.
The number of labels can be adjusted using the default setting segmenting the data into “low”, “medium” and “high” labels. The discretization can be done with different methods. The first method is the division of the data into segments of the same range or into segments including the same amount of data. The Gaussian mixture model is a more sophisticated approach. In case of the Gaussian mixture model the number of labels is chosen by investigating the distribution of the data. For this method the number of maxima in the distribution could be used to indicate the number of segments to be made. The choice of the discretization method mainly depends on the amount of data. If the time window is small only the simple discretization method is applicable.

7.2.2. Finding successive steps

The univariate discretized time series from the first step or additional discrete time series (e.g. wind direction data with cardinal direction labels, or sequence data obtained from earlier experiments) serve as the input. The univariate data consists of time steps with a given label. Here the time series is compressed into triples containing consecutive occurrences of the same label (this is shown Figure 7.1 as Step 1). These triples are defined by their start point, duration of a sequence of the same labels and the label. This data is often interrupted by smaller triples because of the noise in the time series. Smaller triples with a duration shorter than a predefined limit are removed or renamed if this triple is located between two triples of the same label. Two parameters must be adjusted. The first one defines the minimum length of duration for a triple (d_{max}) and the second one controls the size of the gap which is allowed compared to the length of the surrounding segments (r_{max}). For example, a value of r_{max} of 0.5 allows the gap to be half as long as the two surrounding segments together. If this condition is fulfilled the gap is renamed obtaining the name of the surrounding triples.

7.2.3. Merging the discretized data

This step merges the univariate triple into one matrix defining the start point, duration and the combination labels of all given univariate time series. Every time step, a label changes in one of the univariate time series a new combined label is defined (this is shown Figure 7.1 as Step 2). After merging the univariate time series, the filter removing small interruptions can be applied a second time. For example, as a result of this merging the following triple could be created such as (start: 21.07.2015 15:30:00; duration: 60 minutes; label: high temperature, low humidity, hlow wind speed). This means that for the given date and time the three univariate time series were given the mentioned labels lasting for 60 minutes. Subsequently, the labels in one or more univariate time series changes such that a new triple is created.

7.2.4. Finding sequences

In the last step sequences are searched from the merged triples in the previous step (this is shown Figure 7.1 as Step 3). Therefore, a time window must be defined. The time window can range from several hours to days. Here we use always a time window of 1 day, as we are particularly interested in the diel behaviour of trace gases at the measurement locations. Some prior knowledge of the sought phenomena must be brought in. We applied the SPADE algorithm for mining the sequences.[167] This algorithm allows the definition of a gap controlling the amount of omitted triples in a sequence.

7.2.5. Labelling of the time series data

The extracted sequences can be used to label the time series data. For each mined sequence a new variable is added. This variable is of the same length as the original time series and contains the information for each time step if it is included in the sequence ("sequence data") or not ("non-sequence data"). These labelled time series serve as a starting point for further analysis. This includes the application of supervised classification models to identify trace gases which behave differently between "sequence data" and "non-sequence data". Similarly, two sequences can be compared.
Secondly, this labelling allows the extraction of data out of the time series belonging to a specific sequence. This bears the advantage that the behaviour of VOCs for specific meteorological conditions can be chosen (including meteorological variables in the sequence mining step). For example, hierarchical clustering can be used to divide the VOCs into groups of similar behaviour. The comparison between two clustering results of a different sequence allows identification of VOCs behaving differently in these two regimes. Additionally, the affiliation of a VOC to a group or to another VOC was examined using a randomForest model.

Hierarchical clustering

Using this unsupervised method, the time points included in the chosen sequence must be extracted. This must be done for all desired VOCs. Then the extracted data sequences are scaled to make them comparable to each other. From the scaled data a distance matrix is constructed by calculating the Euclidean distance from each pair of the VOCs. Finally, this matrix is used to apply the unsupervised clustering method. Here we used a hierarchical clustering approach. The complete agglomeration method was applied. From the resulting dendrogram different groups of VOCs can be distinguished.
Additionally, it is possible to compare two clustering results. This can be done visually by contrasting the two dendrograms and charting the ways where the VOCs cluster into a different position. This pairwise comparison can be done mathematically calculating the similarity of two dendrograms. This can be used to identify sequences which behave similarly to the way VOCs cluster into the same groups.

RandomForest model to predict the affiliation of an unknown mass

Secondly, to explore the nature of the unknown masses, a model with involving domain knowledge can be built. Here the instances contain all or a subset of the known masses. Then one desired sequence is chosen and the data for all given compounds is extracted and scaled. This is important since the behaviour of the classes is compared to the unknown masses. Finally, a random forest model is built based on this data with all chosen masses. This model is used to predict the mass labels for the instances of the unknown masses. The final table contains the unknown masses and the occurrences of each label given to them. This table can be compared to a second sequence to detect VOCs which change their behaviour.

7.2.6. Regression tree model to estimate the influence of meteorological variables on the measured VOCs

The identified sequences can be used to estimate the influence of the meteorological variables on the measured VOCs. In such cases, two sequences can be chosen along with the desired trace gas. The time series of the VOC and the meteorological variables for each sequence are averaged over the time. This results in one averaged time series ranging for example for sequence 38 from 06:00 to 15:00. Then the difference for each variable for the two chosen sequences is calculated. In the case of categorical meteorological variables such as local wind direction or origin of the air mass the most frequent category for each time step is chosen. This approach results in one variable for numeric variables (difference between the time series of the two sequences) and in two variables for categorical variables (one separate variable for each sequence). Here we choose to apply a regression tree model because of their easy visualization and interpretation and its capability of incorporating non-linear relationships.

7.3. Results

First, the results of pattern identification algorithm for the CYPHEX data are discussed. In Figure 7.2 the times series of temperature, humidity and wind speed are shown. The parameters give to the algorithm were $d_{max} = 3$, $r_{max} = 0.3$, $min_d = 3$, support = 0.3 and maxgap = 7. Additionally, two different sequences are shown. In total, the pattern identification algorithm for the CYPHEX data including temperature, humidity and wind speed resulted in 43 found sequences.

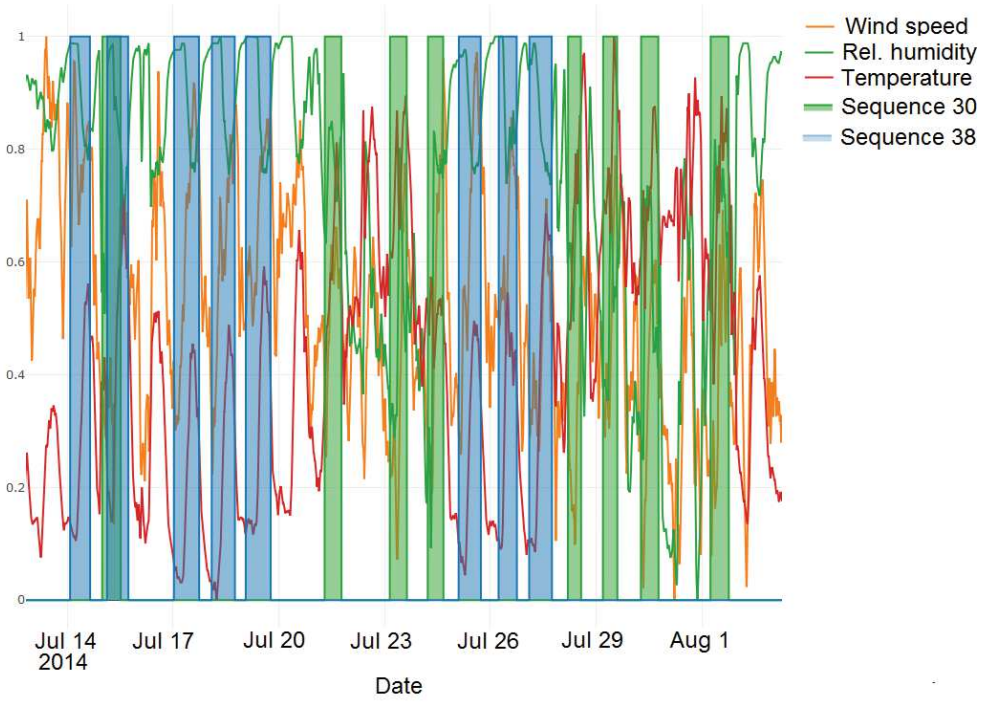


Figure 7.2.: Temporal patterns found using temperature (red line), relative humidity (green line) and wind speed (yellow line). Sequence 30 is shown within the green boxes and sequence 38 within the blue boxes.

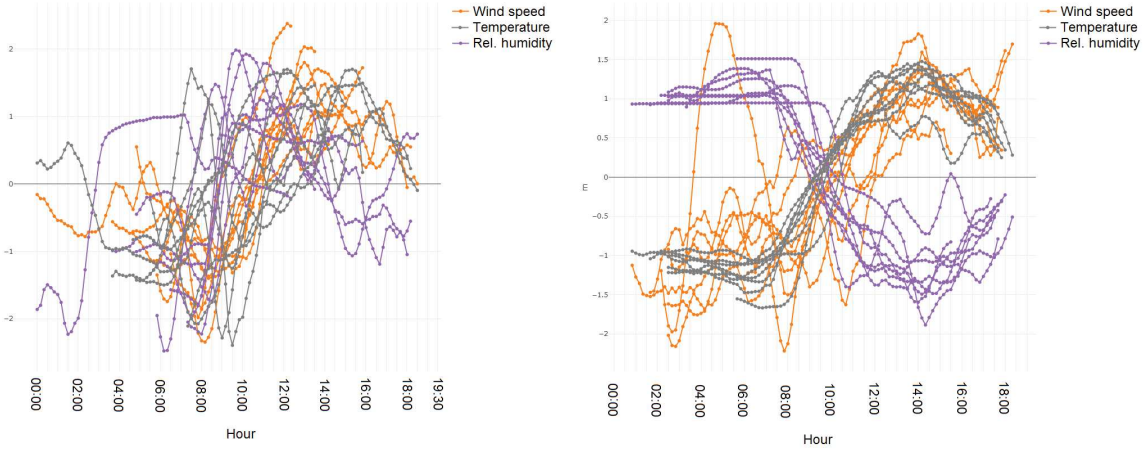
Interesting patterns can be found by examining the fraction of the occurrence of one sequence divided by the number of total time windows (here days) or the total coverage of the sequences in the whole measurement time period. For example, sequences of the length one (for example sequence 11 in Table 7.2), containing only one triple, mostly occur in all time windows but cover only a small amount of the whole time period. This would result in a large fraction (close to 100%) but a small coverage. The results are shown in Table 7.2, including the calculated fraction and coverage of the sequences 30 and 38.

Table 7.2.: Fraction and Coverage for sequences 11, 30 and 38.

Sequence	Fraction	Coverage
11	74%	8%
30	35%	20%
38	35%	25%

The first 16 sequences are of length one, containing only one triple of the combined univariate, discretized variables. The other sequences contain between two to four triples.

The two chosen sequences (sequence 30 and sequence 38) shown in Figure 7.2 are characterized by a fraction of 35% for sequence 30 and sequence 38 and a coverage of 20% for sequence 30 and 25% for sequence 38. For one day (15th July) there is an overlap of both sequences. Other combinations like temperature, humidity and isoprene or ozone result in similar sequences.



(a) Sequence 30.

(b) Sequence 38.

Figure 7.3.: Individual time series for temperature (grey), wind speed (orange) and relative humidity (purple) for sequence 30 (left) and sequence 38 (right).

Figure 7.3 shows the individual time series of temperature, humidity and wind speed for the two chosen sequences 30 and 38. The presented time series data is down-sampled from 1 minute to 10 minutes for the sake of clarity. It can be seen for sequence 38 (on the right side of Figure 7.3) that for one day the wind speed deviates in the time range from 03:00 until 07:00 from the other days. Indeed, the rest of this day behaves like the others showing that this method is robust to outliers and partial deviations. The corresponding written sequences can be found in Table 7.3. For example sequence 38 shows an increase in temperature over the course of day described as a sequence of “low”, “low”, “medium”,“high”.

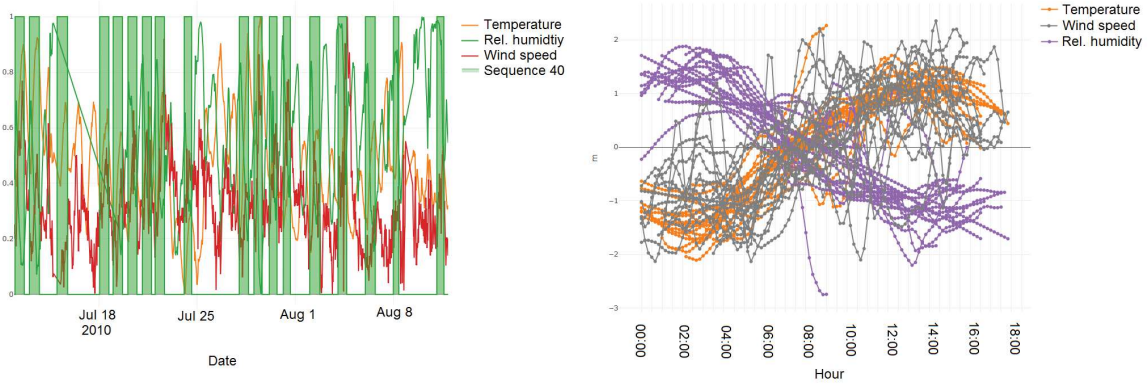
Table 7.3.: Written labels for sequence 30 and sequence 38.

Merged	Sequence 30			Sequence 38		
triples	Temperature	Humidity	Speed	Temperature	Humidity	Speed
1	Medium	Low	Low	Low	High	Low
2	High	Medium	High	Low	High	Medium
3	-	-	-	Medium	Medium	High
4	-	-	-	High	Low	High

For sequence 30, two blocks build this sequence and sequence 38 inherits four blocks. This shows the influence of the maxgap parameter which was set to 7. In case of sequence 30 the sequence mining algorithm only found similar sequences of triples if most of the triples between these two triples were omitted. It is possible to merge two sequences manually if the user decides from the visual investigation that they are similar. This makes sense since the discretization into several segments is arbitrary and depends on the daily behaviour.
The addition of dimethylsulfide, a molecule with a strong marine source,[19] to the meteorological variables leads to a sequence identifying the onset of sea breeze at this island site. Figure 7.4 shows the average behaviour of these variables.

Figure 7.4.: Average behaviour of temperature (brown), relative humidity (yellow), wind speed (green) and dimethylsulfide (orange) during a sequence depicting the onset of the sea breeze.

In order to further examine the potential for data-mining such atmospheric datasets, the meteorological parameters for the HUMPPA-COPEC campaign were used. For this campaign 44 sequences were found. Sequence 40 is shown in Figure 7.5 (left side). This sequence was present for 18 days with a total duration of 31 days. For this data the same daily profile as for sequence 38 for the CYPHEX campaign was extracted (similar daily behaviour of temperature, relative humidity and wind speed). For the rest of the days no distinct sequences were found.



(a) Occurrences of sequence 40.

(b) Individual behaviour during sequence 40

Figure 7.5.: Results of the HUMPPA-COPEC data. On the left side the temporal patterns found using temperature (orange line), relative humidity (green line) and wind speed (red line). Sequence 40 is shown within the green boxes. On the right side the individual behaviour of sequence 40 for temperature (orange), relative humidity (purple) and wind speed (grey) is shown.

7.3.1. Detection of similar behaviour of VOCs and their comparison

Unsupervised approach with hierarchical clustering

In the first stage of this analysis, the known masses were chosen to be included into this unsupervised approach. The hierarchical cluster for sequence 38 is shown in Figure 7.6.

It can be seen that the primarily emitted biogenic compounds (isoprene, pinene and

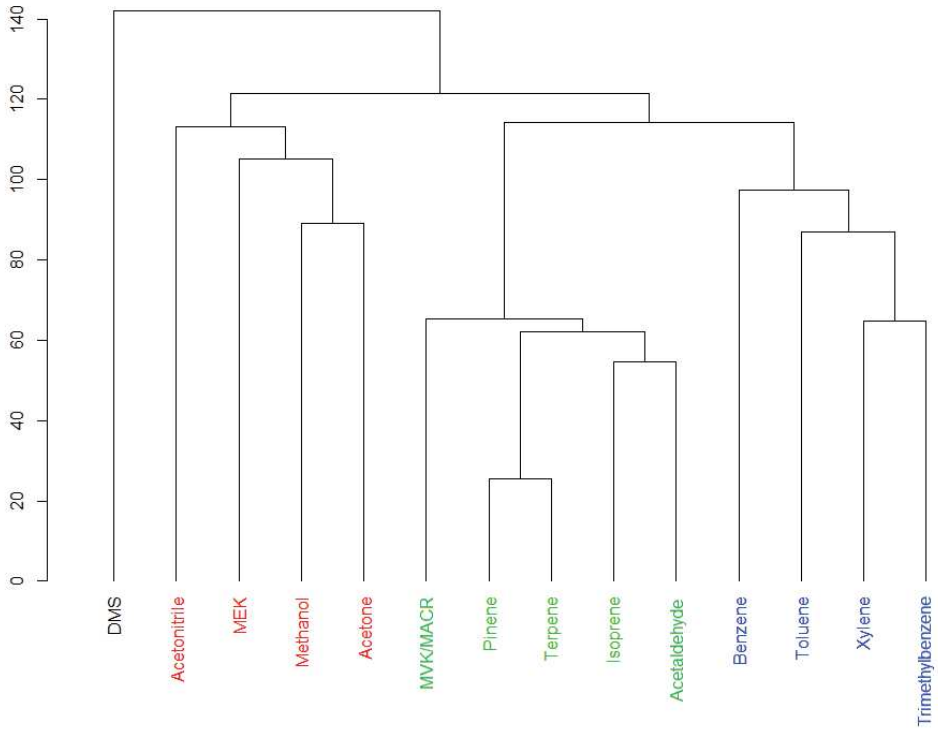


Figure 7.6.: Hierarchical cluster for sequence 38. The colours indicate meaningful clusters comprising anthropogenic compounds (blue), biogenic compounds (green), a red cluster (with oxidized species and acetonitrile) and DMS in black. All known masses were included in this clustering.

monoterpenes) form a cluster together with acetaldehyde and methyl vinyl ketone. A second cluster comprises the primarily emitted anthropogenic compounds including benzene, toluene, xylene and trimethylbenzene. The third cluster is formed by some oxidised compounds (methanol, acetone, methyl ethyl ketone) and acetonitrile. Dimethylsulfide shows no cluster affiliation.

It is possible to compare two dendrograms visually as shown in Figure 7.7. The left dendrogram shows the hierarchical clustering result from sequences 30 and the right one from sequence 38. It can be seen that mainly dimethylsulfide (from "biogenic" cluster in sequence 30 to no cluster affiliation in sequence 38), acetaldehyde (from "anthropogenic" cluster in sequence 30 to "biogenic" cluster in sequence 38) and methyl vinyl

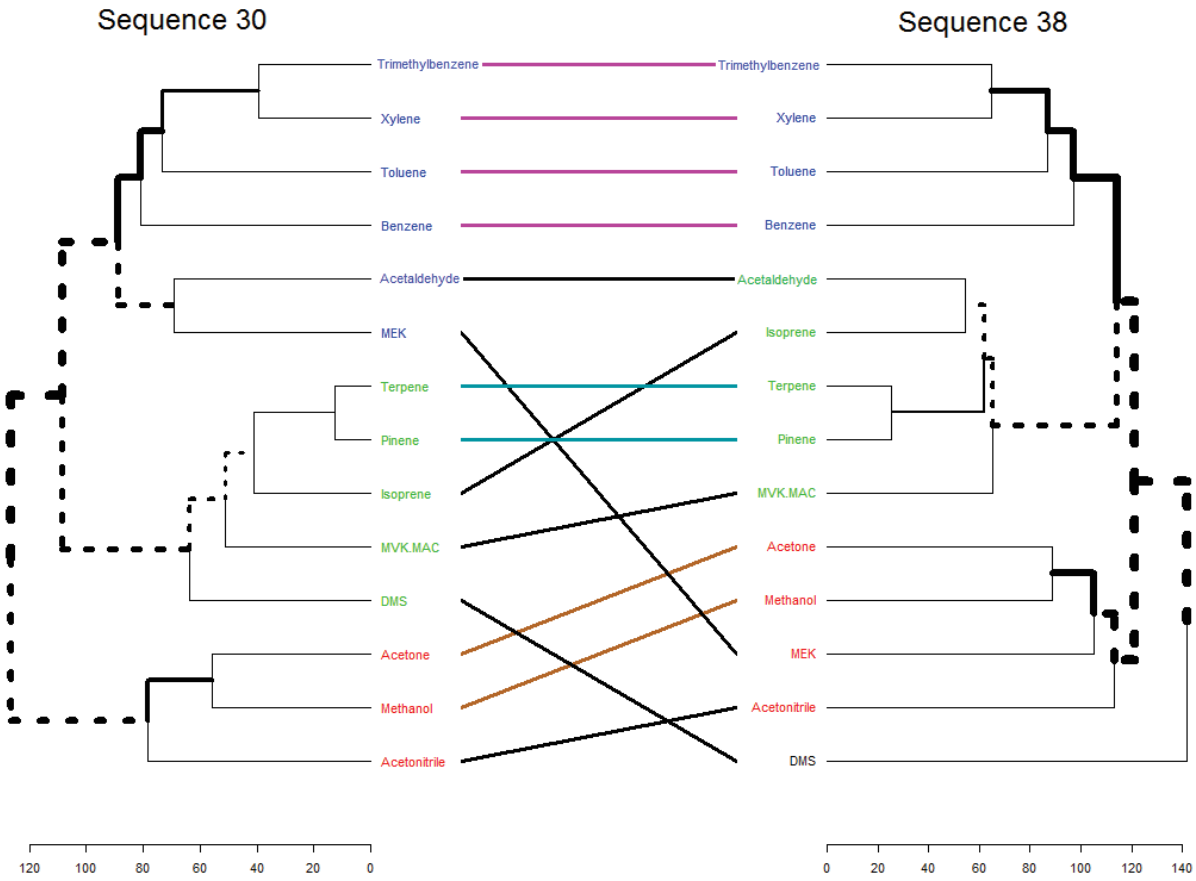


Figure 7.7.: Comparison of the hierarchical cluster from sequence 30 (left) to the cluster for sequence 38 (right).All known masses were included for this comparison.

ketone (from "anthropogenic" cluster in sequence 30 to "oxidized" cluster in sequence 38) change their cluster affiliation. Here is to say that the clusters are chosen according to domain knowledge and there is no rule whether these clusters should be formed with a finer (more accurate) or coarser separation. The height of the knots depicts the Euclidean distance between two VOCs or clusters. Thus the time series of monoterpenes and pinene are more similar than the time series of xylene and trimethylbenzene. For acetonitrile it can be seen that the similarity is small and the affiliation to one cluster is remote. A cluster affiliation is not always unambiguous and can vary with different distance metrics and linkage criteria.

Figure 7.8 shows the comparison of two dendrograms with 40 masses. The comparison of these two dendrograms becomes ambiguous and it gets hard to pick any clusters or follow changes in cluster affiliations of VOCs between the sequences. Therefore, we applied a supervised approach with a randomForest model.

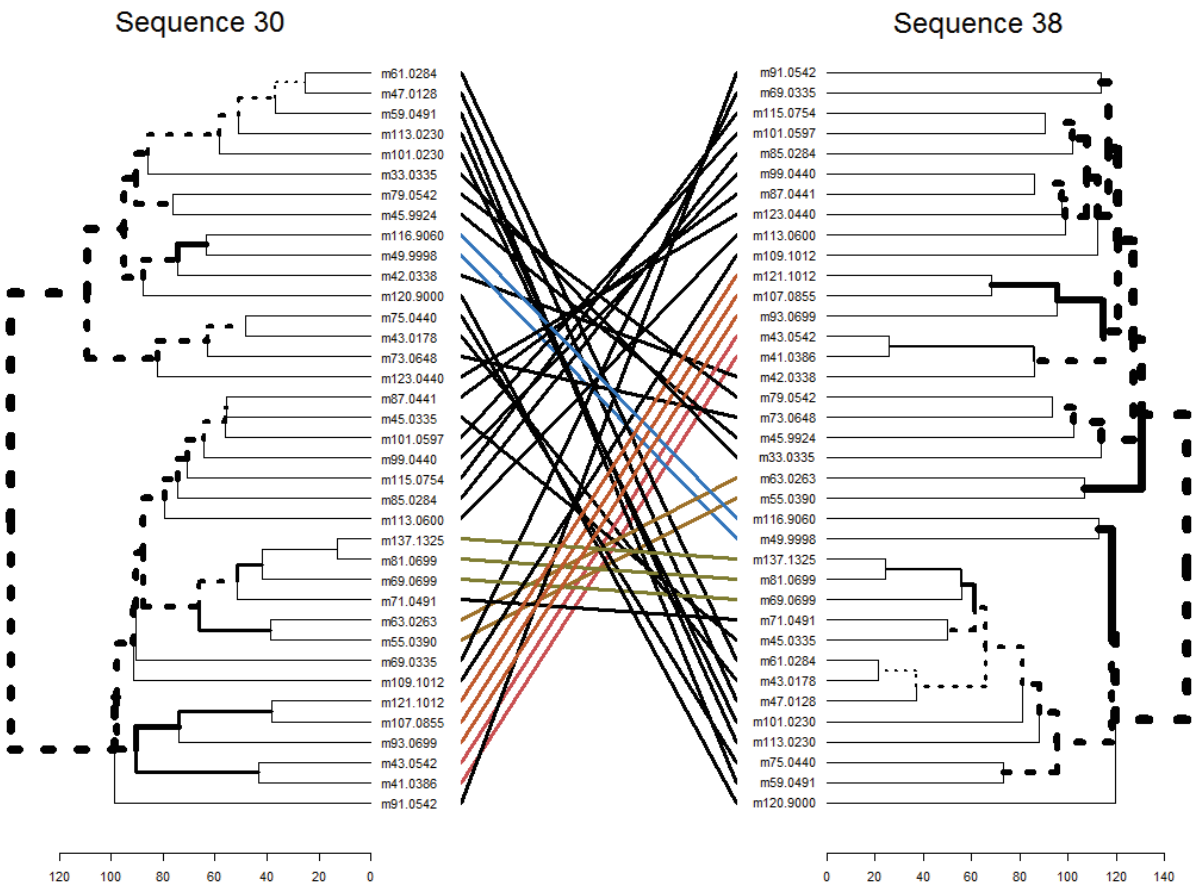


Figure 7.8.: Comparison of the hierarchical cluster for sequence 30 (left) to the cluster for sequence 38 (right) for all measured masses.

Supervised approach with a randomForest model

The results from the supervised approach with a randomForest model yields a matrix with the number of instances of the unknown masses which were classified as one of the known masses. An example is shown in Table 7.4.

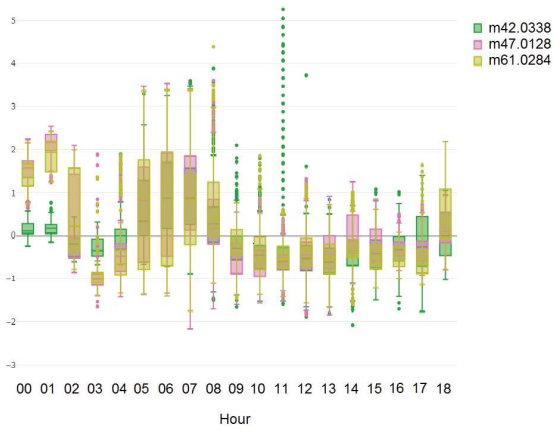
Figure 7.9 shows the results for ethanol (m47.0128) and acetic acid (m61.0284). In

Table 7.4.: Example from the randomForest approach showing three unknown masses and the occurrence how often they were classified as one as the known masses.

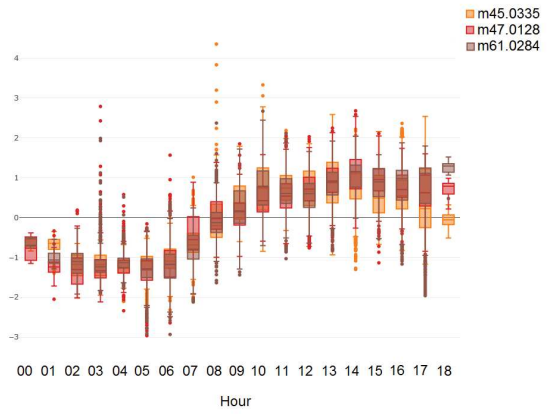
	Sequence 30			sequence 38		
mass	Acetaldehyde	Acetonitrile	Benzene	Acetaldehyde	Acetonitrile	Benzene
m47.0128	79	223	33	138	0	8
m61.0284	85	207	15	160	0	41
m113.0230	52	74	196	17	17	100

sequence 30 these masses are correlated with acetonitrile. During this sequence acetaldehyde clearly shows a different pattern. During sequence 38 ethanol and acetic acid

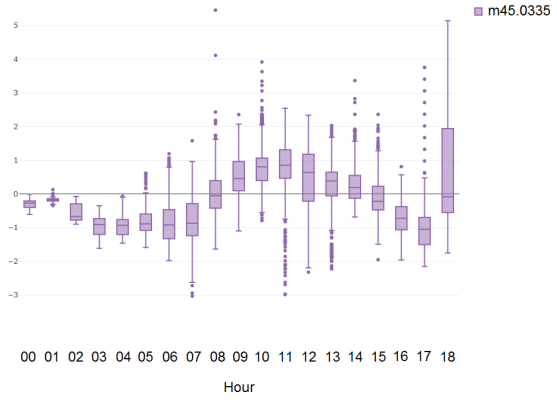
are correlated with acetaldehyde and acetonitrile behaves differently.



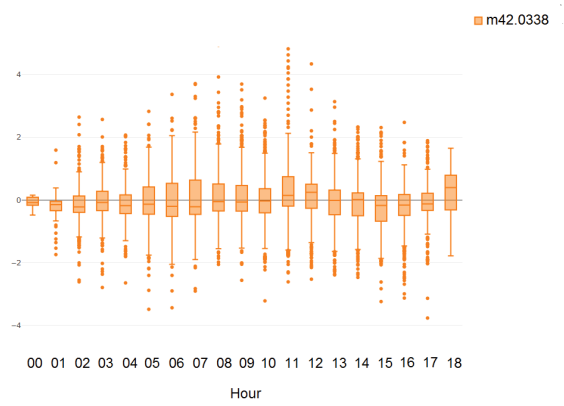
(a) Hourly boxplots of acetonitrile, ethanol and acetic acid during sequence 30.



(b) Hourly boxplots of acetaldehyde, ethanol and acetic acid during sequence 38.



(c) Hourly boxplot of acetaldehyde during sequence 30.

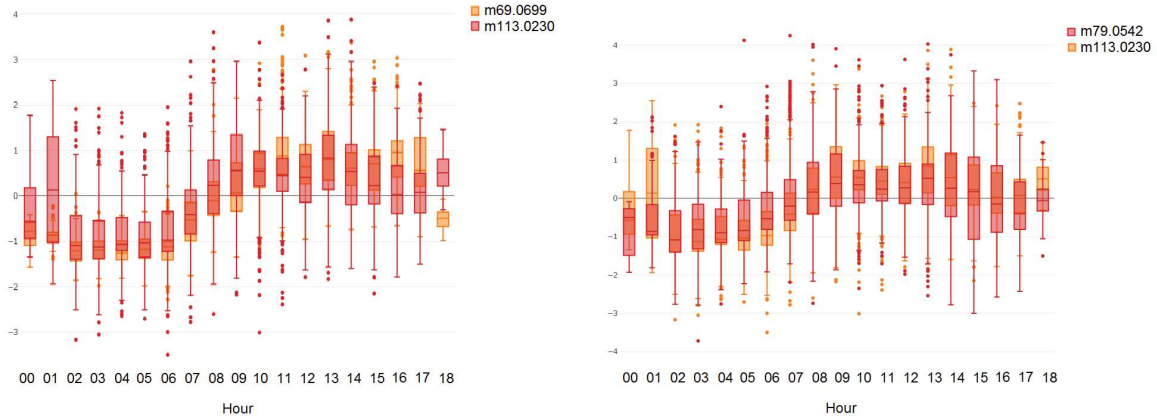


(d) Hourly boxplot of acetonitrile during sequence 38.

Figure 7.9.: On the left side daily behaviour of VOCs for sequence 30 and on the right side for sequence 38 are shown. The left top panel shows the behaviour of acetonitrile (m42.0338), ethanol (m47.0128) and acetic acid (m61.0284). The left bottom panel shows the behaviour of acetaldehyde (m45.0335). The right top panel shows the daily behaviour of acetaldehyde, ethanol and acetic acid and the right bottom panel acetonitrile.

In comparison with the results from the distance-based approach with the hierarchical cluster from the previous section it can be seen that the supervised approach may yield more stable results. This can be seen for the mass m113.0230. In the dendrogram for sequence 30 and sequence 38 this mass forms a cluster together with ethanol and acetone. For sequence 30 close proximity to benzene can be found whereas for sequence 38 isoprene and other biogenic emitted compounds are neighbouring m113.0230. The supervised approach affiliates m113.0230 for both sequences to benzene. Figure 7.10

compares the behaviour of m113.0230 with benzene and isoprene during sequence 38. Benzene and isoprene seem to be both capable of following the behaviour of m113.0230 but benzene seems to lie closer to the daily profile of m113.0230.



(a) Hourly boxplot of isoprene (orange) and m113.0230 (red).

(b) Hourly boxplot of benzene (red) and m113.0230(gray).

Figure 7.10.: Comparison between the behaviour of isoprene (m69.669), benzene (m79.0542) and m113.0230 during sequence 38.

7.3.2. Regression tree model to estimate the influence of meteorological variables on the measured VOCs

Estimating the influence of meteorological variables on the abundance of VOCs was performed using temperature, relative humidity, wind speed, mixed layer depth (mld), local wind direction and air mass origin as independent variables. It should be noted that the variables temperature, relative humidity and mixed layer depth are highly correlated (correlation factors: $r_{\text{humidity, temperature}}$ = -0.92, $r_{\text{humidity, mld}}$ = -0.93,$r_{\text{temperature, mld}}$ = 0.87). This multicollinearity may pose a problem to the reliable modelling of the regression trees since small changes in one of the variables can alter the representation of the whole tree. The most abundant wind direction and air mass origin are shown in Table 7.5.

Table 7.5.: Summary of the wind direction and air mass origin (cluster affiliation). The numbers indicate the proportion of the abundance for each level.

Local wind direction		Air mass origin (cluster affiliation)	
SW	0.46	Cluster 5	0.36
WSW	0.21	Cluster 6	0.23
SSW	0.14	Cluster 1	0.12
W	0.05	Cluster 4	0.10
S	0.04	Cluster 3	0.08
WNW	0.03	Cluster 2	0.06
Other	0.06	Other	0.05

In case of the air mass origin, an hierarchical clustering approach on the back trajectories was used to assess the number of clusters. The back trajectories with their cluster affiliation are shown in Figure 7.11. For the clustering of the back trajectories the longitude and latitude as well as the height was included.

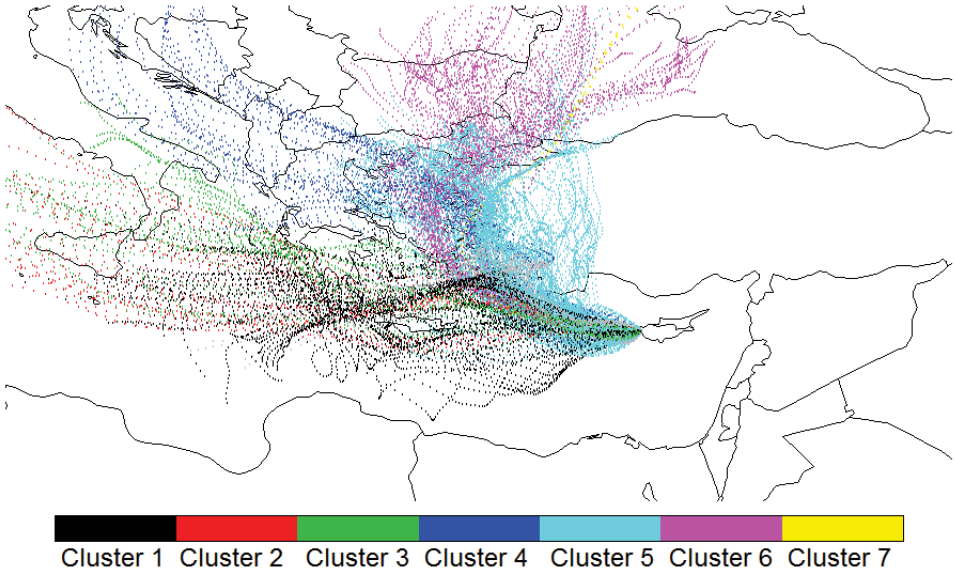
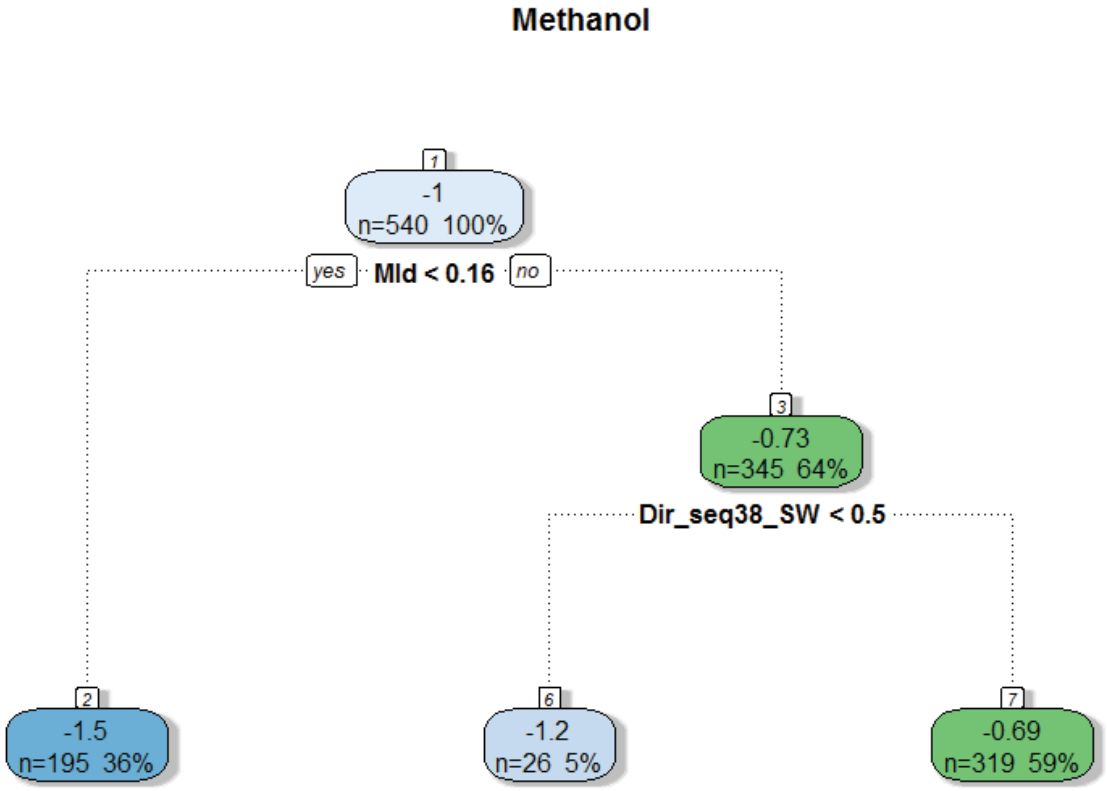


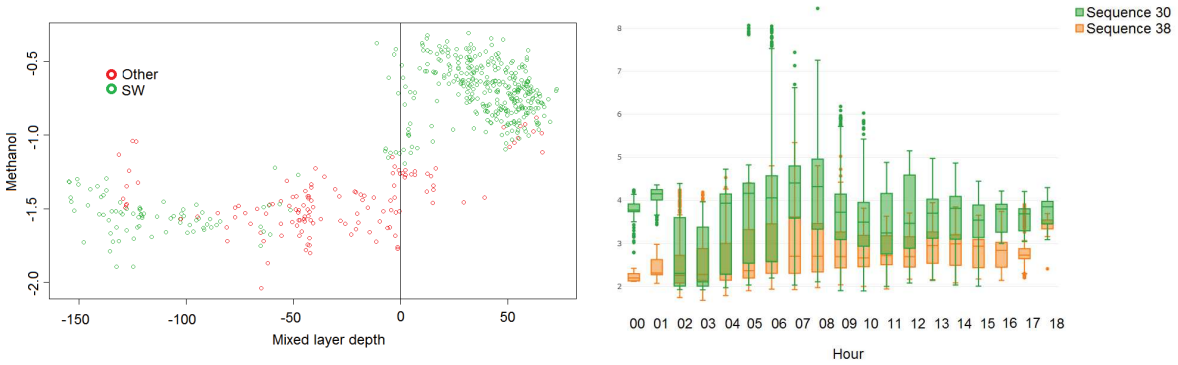
Figure 7.11.: Back trajectories during the whole measurement period. The colours indicated the cluster affiliation.

The results of the regression trees show that compounds like methanol, acetone, acetonitrile, and DMS are mostly linked to humidity. Biogenic compounds such as isoprene and monoterpenes are dependent on temperature and anthropogenic compounds (aromatic compounds) mostly on the air mass origin. An example of the visualization of a

regression tree model is shown in Figure 7.12 for methanol. The top box (box 1) summarizes the average value of the difference of methanol between sequence 38 and 30 of the complete data set (average of -1 indicating overall higher mixing ratios of methanol during sequence 30 than during sequence 38). The boxplot of the daily behaviour during these two sequences is shown in the lower right part in Figure 7.12 showing sequence 30 in green and sequence 38 in orange.
The second line in the boxes of the tree model representation shows the number of instances included in this node and the proportion to the total number of instances. The top box describes the initial condition containing all instances (n = 540) resulting in a proportion of 100%. With a time resolution of 1 minute this model covers 9 hours. Under each non-terminal node the splitting condition is shown. In general, the notation of the splitting nodes is as following: "<variable name><condition>". First the variable name is presented on which variable the split is performed and then a condition which is the splitting criteria. The information gain is used as a statistical property deciding which variable is used for the split. In general the higher the nodes the more important is their influence on the regression. The top node splits the data set into partitions with a difference in the mixed layer depth smaller than 0.16 or greater or equal than 0.16. This split can be seen in the lower left panel in Figure 7.12 indicated by the vertical line. If the humidity is smaller than 0.16 the left path is taken resulting in an average value of -1.5 for methanol. This splitting leads to a terminal node which includes 195 instances (36% of the total data size). If the mld is greater or equal than 0.16 and thus the condition mld < 0.16 is FALSE the right path is taken leading to a non-terminal node. This partition of the data possesses an average value of -0.73 including 64% of the data. For this partition of the data a second split is performed using the wind direction variable of sequence 38. The splitting variable is called "Dir_seq38_SW". This variable is a binary categorical variable containing values of 1 for TRUE and 0 for FALSE. It should be noted that for categorical variables there are separate binary variables for each sequence (sequence 38 or sequence 30) and each category (for example different variables are created for wind directions like "south, "south west", "west" and so on). The binary categorical variables possess only values of 1 or 0. The value 1 indicates that the wind comes from the direction "SW" whereas the value 0 stands for any other wind direction. The notation "Dir_seq38_SW < 0.5" means if the variable "Dir_seq38_SW" has the value of zero, the condition is set TRUE and the left path is taken resulting in a leaf node. If the condition is FALSE meaning that the cluster variable has a value of 1 the data is put into the right path leading to another leaf node. The colour code in the lower left panel in Figure 7.12 shows the division of the data into instances containing the wind direction "SW" (green points) or not (red points). This condition does only apply for the points with a mld larger than 0.16 (located on the left part of the vertical part). Thus for the wind direction "SW" this results in a higher average value (-0.69 for "SW" compared to -1.2 for all other wind directions).



(a) Tree model representation.



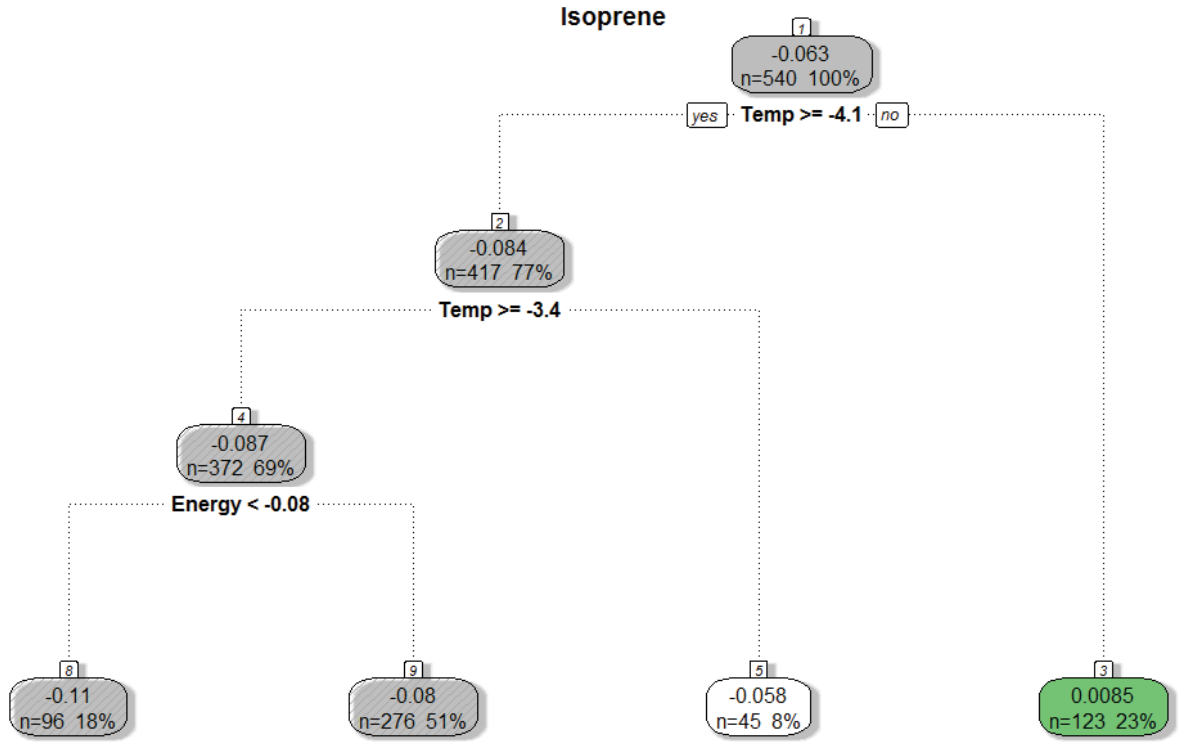
(b) Scatter plot of mixed layer depth and methanol.

(c) Hourly boxplot of methanol during sequence 30 and sequence 38.

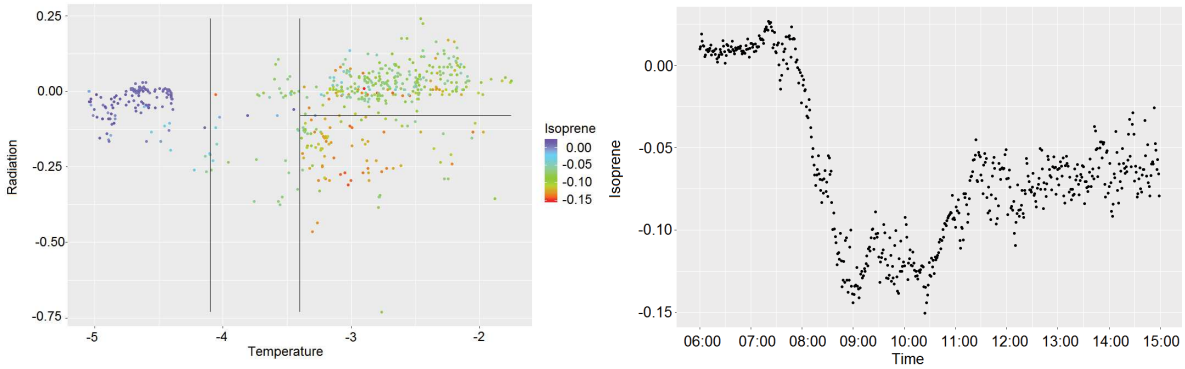
Figure 7.12.: Representation of the regression tree model (top). On the lower left side, the scatter plot of the difference in mixed layer depth versus the difference in methanol between sequence 30 and 38 is shown. The vertical line defines a humidity level of 14 and the green colours depicts instances influenced by a south western (SW) local wind direction during sequence 38 and the red points depict every other wind direction during sequence 38. The boxplot of the daily behaviour of methanol is shown on the lower right (right) for sequence 30 (green) and sequence 38 (orange).

7.4. Discussion

The results presented in the previous section raise several interesting questions about the origin and fate of VOCs in the atmosphere. They were effectively distilled from the data sets using the data mining approaches described above. Therefore, we now evaluate the comparison of the two dendrograms in Figure 7.7 in conjunction with the results from the regression tree model from the perspective of atmospheric chemistry. The comparison of the dendrograms in Figure 7.7 resulted in roughly three groups of molecules. The first group includes biogenic masses like isoprene, pinene, monoterpenes, methyl vinyl ketone/methacrolein (MVK/MACR). The second group includes anthropogenic masses like benzene, toluene, xylenes and trimethylbenzene. The third group includes oxidized species like methanol, acetone as well as dimethylsulfide and acetonitrile.
First we discuss the biogenic compounds comprising isoprene, pinene, monoterpenes and the oxidation product methyl vinyl ketone/methacrolein. These masses cluster together for both sequences in the "biogenic" cluster as shown in Figure 7.7. It is well established that the primary emission of isoprene, pinene and other monoterpenes is primarily driven by temperature and sunlight radiation. The atmospheric lifetime of these molecules ranges from one hour to several hours during daytime.[73, 137] This rather short lifetime leads to the assumption that the measured mixing ratios are mostly influenced by local emissions from vegetation. MVK/MACR is an oxidation product of these primary emitted compounds but has also been reported as a primary emission.[65] Isoprene was chosen as an example for these biogenic compounds. In Figure 7.13 the representation of the regression tree model and its partition with the two variables temperature and energy is shown.
It can be seen in Figure 7.13 that the first two splits use the temperature variable. During sequence 30 a higher temperature is faced resulting in a negative temperature difference (between -5 and -2°C). Between 4:00 and 7:00 the largest temperature difference is calculated whereas during this time the isoprene mixing ratios were almost equal (purple colour coding in Figure 7.13). The isoprene emission is mainly triggered by sun radiation which rises around 07:00.[124, 152] Therefore, no large difference in isoprene mixing ratio between 4:00 and 7:00 is expected. The third node in the tree model uses the sun radiation. Here a higher sun radiation during sequence 30 (more negative difference) leads higher isoprene emissions (more negative isoprene difference) or vice versa.



(a) Tree model representation.



(b) Scatter plot of temperature and sun radiation.

(c) Difference in isoprene mixing ratio between sequence 38 and sequence 30.

Figure 7.13.: The upper part shows the representation of the tree model for isoprene. The lower left panel shows the scatter plot of temperature and sun radiation. The colour scaling indicates the difference in isoprene between sequence 38 and sequence 30 (ranging from purple for a value of ca. 0.00 to red for a value of ca. -0.15). The horizontal and vertical lines represent the partition of the regression tree model. The lower right panel shows the difference of isoprene between sequence 38 and sequence 30.

The other biogenic compounds such as pinene, monoterpenes and MVK/MACR behave similar to isoprene and use the temperature variable for their first splitting. MVK/MACR

then uses the air mass origin variable for the next splitting accounting for the transport of these molecules. MVK/MACR possess an atmospheric lifetime of 6-10 hours during daytime.[46] If the air mass originates from cluster 5 in Figure 7.11 higher mixing ratios during sequence 30 were measured than for cluster 1. Thus the air mass from cluster 5 is maybe stronger influenced by photo-processed air with primary and secondary sources in closer proximity. Whereas the air mass from cluster 1 spent a longer time over the sea with no primary or secondary sources of MVK/MACR. The lower mixing ratio for cluster 1 suggests that either the emission sources where stronger for Eastern Europe or the removal processes had a stronger impact for the air masses originating from cluster 1.[30] Pinene and the monoterpenes also use the temperature variable as the first split and then use the local wind direction variable as the third splitting criteria. If the wind comes from the southwest during sequence 38 higher emissions of these compounds were measured compared to the other directions. Closely in this direction a pine tree-covered area is located which could lead to higher mixing ratios.
The second group includes the anthropogenic compounds such as benzene, toluene, xylenes and trimethylbenzene. These aromatic species form one cluster in the dendrograms presented in Figure 7.7. The representation of the regression tree model shows that for all these masses except for toluene, the most important variable is the "cluster_seq30_clusNum1"-variable (see Figure 7.14. This variable includes the information about the first cluster of the air mass origin for sequence 30. This node means that if the air mass during sequence 30 comes from cluster 1 (condition is FALSE and thus going the right path) the mixing ratios of xylenes for sequence 30 is lower (more positive difference). On the other hand, if the air mass does not come from cluster 1 (condition "cluster_seq30_clusNum1 < 0.5" is TRUE) the mixing ratios during sequence 30 are higher. This makes sense since the other two most abundant clusters are cluster 5 and 6 (see Table 7.5) and these air masses originate from Eastern Europe (Western Turkey see Figure 7.11) which is geographically closer and therefore the air is carries higher mixing ratios of anthropogenic compounds. In contrast, cluster 1 spends most of its time of the Mediterranean Sea (see Figure 7.11) with less or no sources of these aromatic compounds. The presented aromatics are mostly emitted from fuel combustion processes,[3, 22, 68] biomass burning[165] or industrial solvent evaporation.[116] The atmospheric lifetime of toluene is 1-2 days and for benzene around 12 days allowing this compound to be transported from more remote sources to the measurement site.[44] For benzene and trimethylbezene also the "cluster_seq30_clusNum1"-variable for the first split was used. Interestingly, the splits for toluene was performed using the humidity and temperature difference.

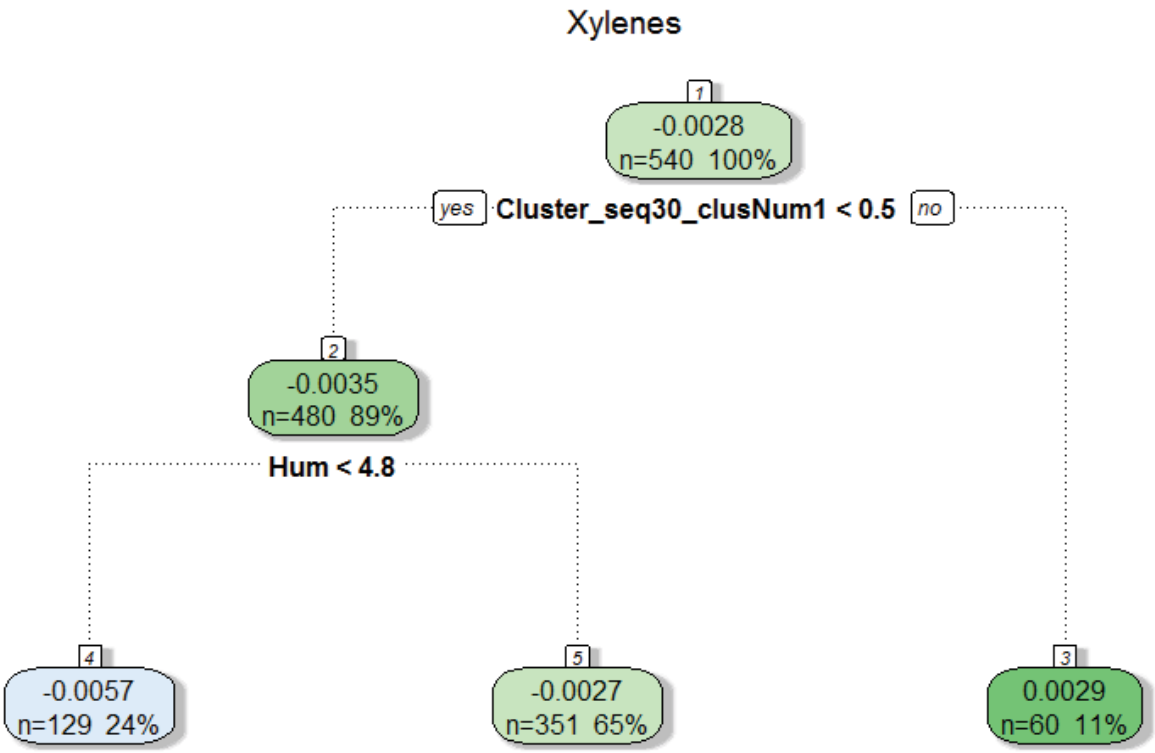


Figure 7.14.: Representation of the tree model for the xylenes.

Figure 7.15 shows the cluster dendrogram for the differences between sequence 38 and sequence 30 for all known VOCs. It can be seen that the behaviour of the difference of toluene resembles the behaviour of the biogenic masses. It is suggested that the abundance of toluene is stronger influenced by the local emission of toluene by the vegetation.[97] Despite the known biochemical production of a range of aromatic compounds by plants and the presence of benzenoids in floral scents, the emissions of only a few benzenoid compounds have been reported from the biosphere to the atmosphere. Here, using evidence from measurements at aircraft, ecosystem, tree, branch and leaf scales, with complementary isotopic labelling experiments, we show that vegetation (leaves, flowers, and phytoplankton) emits a wide variety of benzenoid compounds to the atmosphere at substantial rates. Controlled environment experiments show that plants are able to alter their metabolism to produce and release many benzenoids under stress conditions. The functions of these compounds remain unclear but may be related to chemical communication and protection against stress. The estimatation of the total global secondary organic aerosol potential from biogenic benzenoids is similar to that from anthropogenic benzenoids (ca. 10 Tg y^{-1}), pointing to the importance of these natural emissions in atmospheric physics and chemistry. This was also found by White et al. indicating the contribution of pine trees to the emission of toluene. Pine trees also dominate the vegetation in Cyprus.[150] To summarize, the emission of toluene is clearly dominated by anthropogenic emission for both sequences (shown in Figure 7.7) and smaller differences between the two sequences may occur because of local emissions

from the surrounding vegetation.

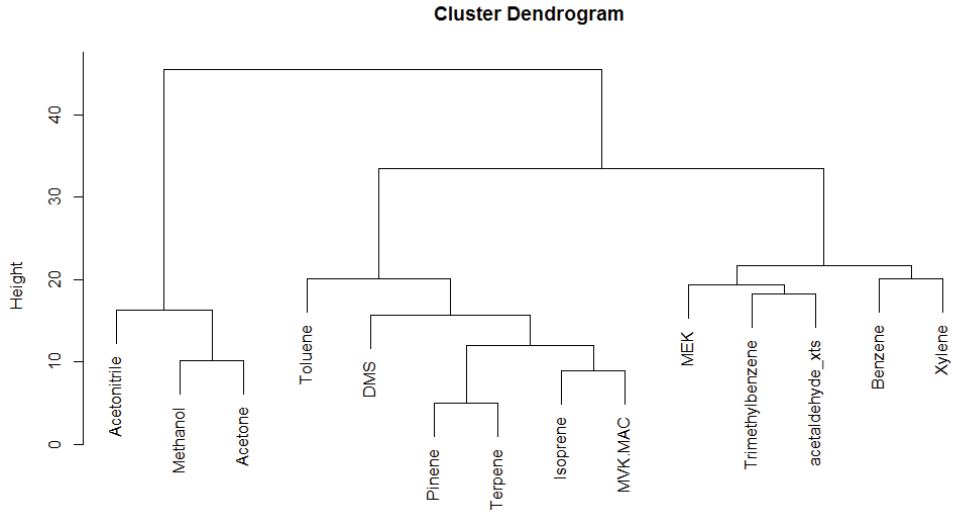
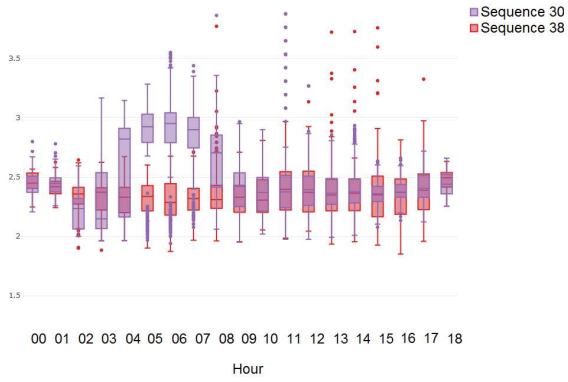


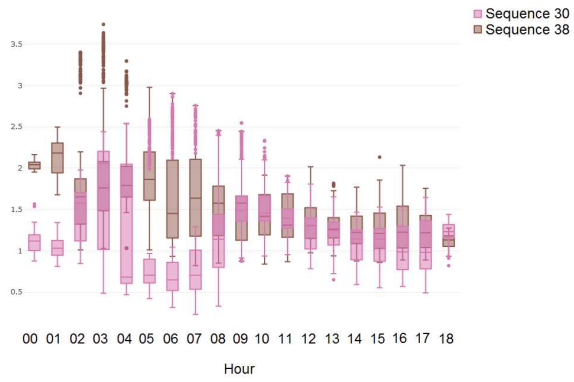
Figure 7.15.: Dendrogram of the hierarchical clustering of the differences between sequence 38 and sequence 30 for all VOCs.

The third group includes compounds like methanol, acetone and acetonitrile. These species also form a cluster in Figure 7.7. They mainly depend on relative humidity according to the regression tree model. Methanol uses as the only examined species the mixed layer depth as its first splitting criteria. However due to its water solubility and the similarity to acetone and acetonitrile in Figure 7.7 and Figure 7.15 we will discuss these species together. Additionally, the correlation between the humidity and mixed layer depth is highest for all combinations ($r_{\text{humidity, mld}}$ = -0.93). Furthermore, using mld as the splitting criteria for methanol does not make much sense since this splitting node states that if the difference in mld is smaller than 0.16 the higher the mixing ratio for methanol in sequence 30 compared to sequence 38 (more negative values for the difference in methanol). This means that the if the mixed layer depth becomes higher the larger the mixing ratio gets (more negative differences for mld and methanol). This is in contrast to the physical understanding because the higher the mixed layer depth the lower the mixing ratio due to the larger turbulent mixing of the compound. The tree models for acetone and acetonitrile show that the higher the difference in humidity the lower the mixing ratios of these compounds for sequence 30. Looking at the difference of these compounds between sequence 38 and sequence 30, lower values (meaning higher abundance in sequence 30 and lower values in sequence 38) occur from 04:00 to 09:00 (see Figure 7.16 for acetonitrile). During this time the humidity has its largest difference (higher values for sequence 38 and lower values for sequence 30 in Figure 7.16). Thus

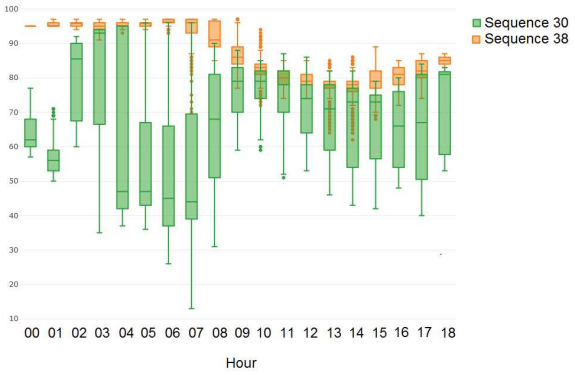
the lower mixing ratios for these compounds maybe stem from their high water solubility. The drop in humidity in sequence 30 may come from the fact that the air mass originates from higher altitudes (free troposphere)[30] and has been more isolated from contact with the water surface than the air masses from sequence 38. Thus these soluble compounds are removed by a lesser amount. The representation of the regression tree of acetonitrile shows the relative humidity as the first splitting node with a second node using the mixed layer depth. The mld is highly correlated with humidity and we assume that acetonitrile is mainly affected by the difference in humidity. We suggest that this comes from the high water solubility of acetonitrile and thus the higher marine influence and increased uptake[28, 53] rather than abundance through biomass burning. Whereas for acetone and methanol an additional node is added. For acetone the air mass origin and for methanol the local wind direction is included into the tree model. This tends to support the assumption made by Derstroff et al.[30] of varying emissions of acetone and methanol due to the source region. Derstroff et al.[30] reported that for the acetone and methanol, the uptake to the sea surface cannot solely defined by solubility and that there is a potential sink in the Mediterranean Sea or emission variability in the source region. The emission variability of methanol and the additional nodes in the regression tree might come from these local biogenic emissions released by the surrounding forests. Acetone possesses a rather long atmospheric lifetime of 1 month (compared to methanol with an atmospheric lifetime of 10 days)[120] and might be stronger influenced by the air mass origin through long-range transport. Interestingly, even though acetone and methanol are known to be emitted from forests,[55, 64, 78] both species are not grouped together with the other biogenic compounds (isoprene, monoterpenes).
For dimethylsulfide the first splitting node contains the humidity variable. It shows the higher the difference in humidity (higher humidity in sequence 38 compared to sequence 30) the higher the mixing ratio of DMS (see Figure 7.16). Since dimethylsulfide is emitted from marine phytoplankton[19] it is obvious that the air mass needs to be in contact with water surface to transport this compound to the measurement site. This fits the behaviour of dimethylsulfide which shows a drop between 04:00 to 09:00 for sequence 30 when the difference in humidity shows the highest values due to the drop of relative humidity during this time for sequence 30. Similar mixing ratios between the two sequences are observed for the rest of the day. It is known that DMS is emitted from vegetation, soil and marshes.[72, 74, 85] The presented analysis does not show that these sources play an important role in this location.



(a) Hourly boxplot for acetontrile.



(b) Hourly boxplot for DMS.

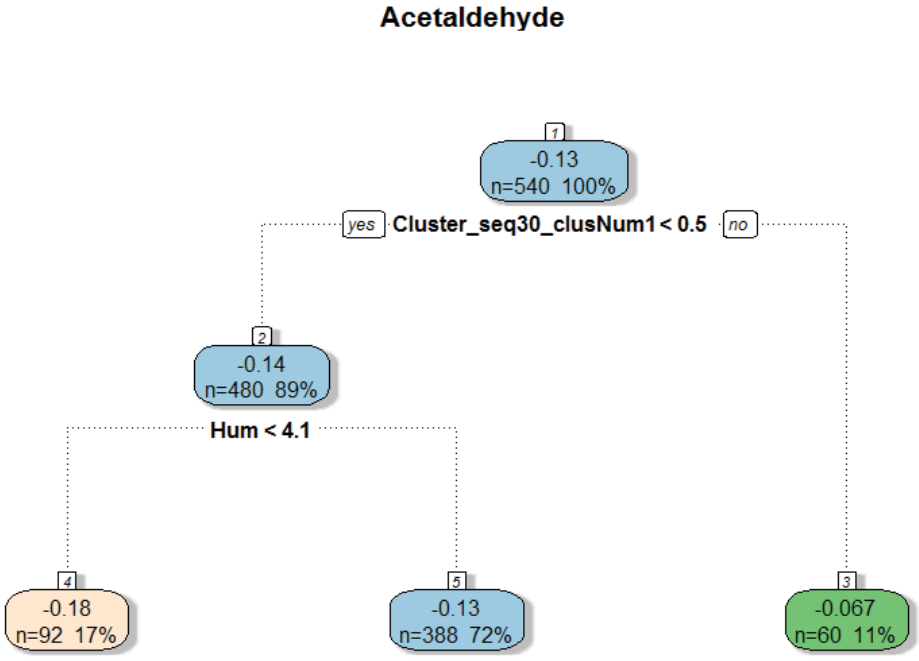


(c) Hourly boxplot for rel. humidity.

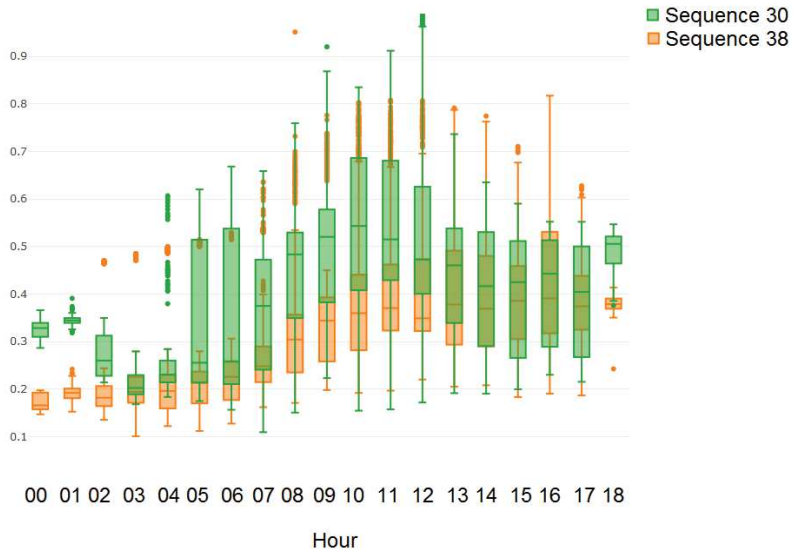
Figure 7.16.: Hourly boxplots for acetonitrile, DMS and rel. humidity for sequence 30 and sequence 38.

The remaining compounds for discussion are methyl ethyl ketone and acetaldehyde. These two compounds can be emitted from anthropogenic and biogenic sources and can be formed through photochemical processes. Acetaldehyde can be formed through the oxidation of multiple alkanes, alkenes, ethanol and also from isoprene the dominant biogenic molecule.[96] Methyl ethyl ketone is an oxidation product of n-butane, 2-butanol, 3-methyl pentane and 2-methyl-1-butene. The photo oxidation of the dominate biogenic VOCs like isoprene and pinene presumably does not yield methyl ethyl ketone.[163] Additionally, there is no linkage to the biogenic cluster (see Figure 7.7 and Figure 7.15). From the comparison of the hierarchical clustering in Figure 7.7 it can be seen that for sequence 30 with higher abundance of anthropogenic compounds MEK and acetaldehyde both cluster together with these anthropogenic compounds. This similar behaviour can be caused by primary anthropogenic emissions of acetaldehyde and MEK or due to the oxidation of anthropogenic precursor molecules. During sequence 38 acetaldehyde forms a cluster with the biogenic compounds probably due to the photo oxidation of these compounds whereas MEK lacking of a pathway for MEK production through dominant biogenic molecules and resembles the behaviour acetone and methanol. However, this cluster including MEK, acetone, methanol, acetonitrile has high x-axis values compared

to the other cluster meaning that these compounds resemble each other only remotely. The representation of the regression tree model for acetaldehyde is shown in Figure 7.17.



(a) Tree model representation.



(b) Hourly boxplot of for sequence 30 and sequence 38.

Figure 7.17.: Representation of the regression tree model for acetaldehyde and the boxplot of its daily behaviour.

It can be seen that the first splitting node for acetaldehyde is the air mass origin during sequence 30. It is the same splitting as for the anthropogenic compounds. Thus it can be guessed that the acetaldehyde production during sequence 30 is strongly influenced by anthropogenic precursor. For sequence 38 with higher proportion of biogenic molecules the behaviour resembles its biogenic precursor molecules. In Figure 7.17 it can be seen the daily boxplot for sequence 38 in orange and sequence 30 in green. For sequence 30 the mixing ratio is similar during the morning and evening hours and are higher during midday. The regression tree model suggests that this difference comes from the different air mass origin. Since the anthropogenic masses shows the same splitting criteria this might come from the higher abundance of anthropogenic molecule which are photo-chemically degraded to acetaldehyde during this time period. This anthropogenic contribution adds on the local primary and secondary production of acetaldehyde. The second node uses the difference in humidity. Indeed, during 04:00 and 07:00 enhanced mixing ratios for acetaldehyde for some days during sequence 30 albeit smaller than for acetone, methanol or acetonitrile can be seen. The tree representation of MEK becomes difficult to explain. The tree model for MEK uses the humidity variable, but in a different way than acetone, acetonitrile and methanol, and the local wind direction variable as splitting criteria.

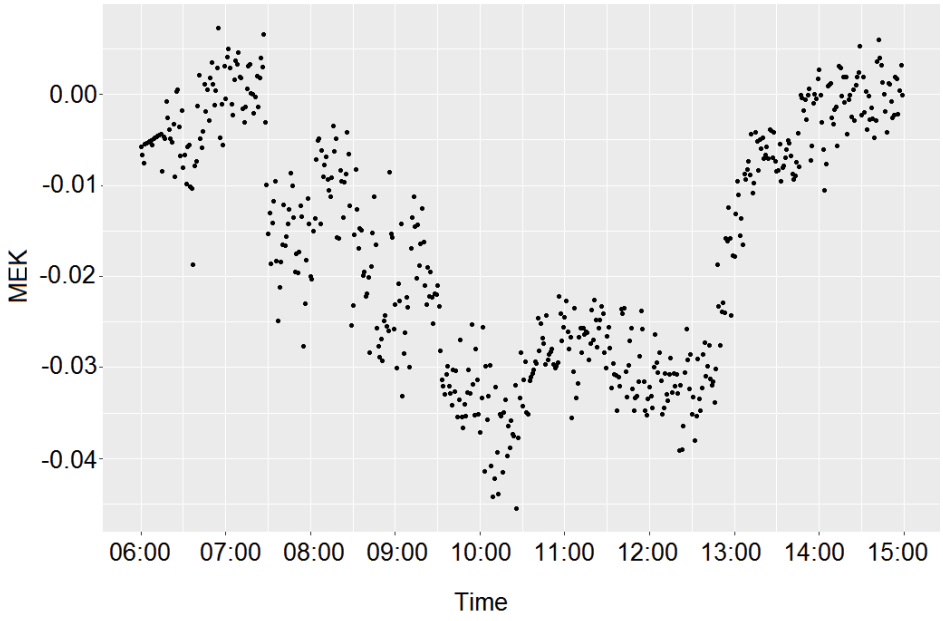


Figure 7.18.: Difference in MEK mixing ratio between sequence 38 and sequence 30.

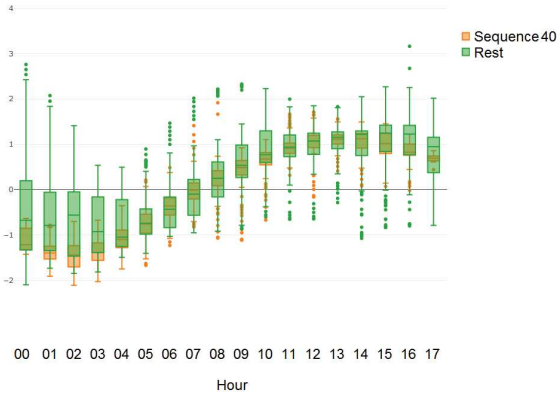
The differenced time series in Figure 7.18 shows a strong minimum during midday. This might be due to the photochemical production of MEK during sequence 30 from the anthropogenic precursor molecules whereas during sequence 38 only lower mixing ratios due to lower anthropogenic precursor molecules were measured.
With this interpretation of the origin and fate of the presented molecules unknown

masses can be affiliated to the known ones. An initial approach might be to include all measured masses in the hierarchical clustering approach. This results in fairly ambiguous results as shown in Figure 7.8. A better approach might be the use of a randomForest model as explained in the previous section.
Table 7.6 shows the results and summarizes the majority mass for each unknown mass from the supervised randomForest approach. It can be seen that most of the unknown masses were predicted as one of the anthropogenic masses.

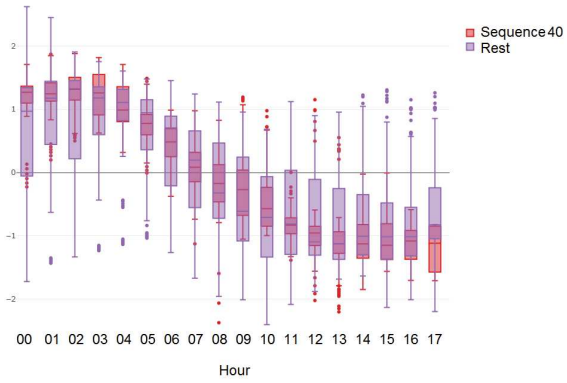
Table 7.6.: Summary of the majority mass from the randomForest approach.

Masses	Sequence 30	Sequence 38
m101.0230	Acetonitrile	Methanol
m101.0597	Toluene	Trimethylbenzene
m109.1012	Benzene	Trimethylbenzene
m113.0230	Benzene	Benzene
m113.0600	Toluene	Trimethylbenzene
m115.0754	Toluene	Acetonitrile
m123.0440	Benzene	Acetonitrile
m41.0386	Toluene	Trimethylbenzene
m43.0178	Benzene	Acetaldehyde
m43.0542	Xylenes	Trimethylbenzene
m47.0128	Acetonitrile	Acetone
m61.0284	Acetonitrile	Acetone
m69.0335	Toluene	Trimethylbenzene
m75.0440	Benzene	Acetone
m79.0542	Benzene	Benzene
m85.0284	Toluene	Trimethylbenzene
m87.0441	Acetaldehyde	Benzene
m93.0699	Toluene	Toluene
m99.0440	Acetaldehyde	Trimethylbenzene

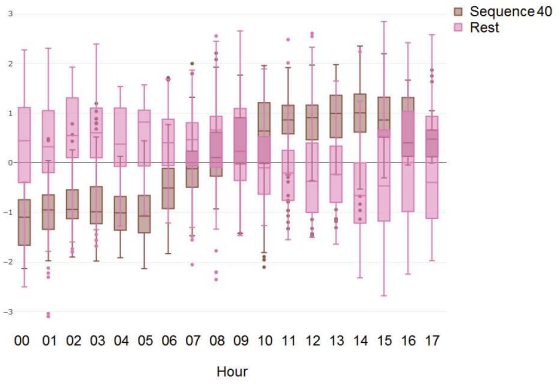
Two unknown masses m109.1012 and m113.0230 had a clear affiliation to anthropogenic compounds. A possible sum formula for m109.1012 is C_8H_{12}. This could 1,5-cyclooctadiene which is commonly used as a ligand in organometallic chemistry and 10.00 tons per year were approximately produced. The second mass m113.0203 could be connected to the sum formula of $C_5H_4O_3$. This could be 2-furoic acid. It is used in food sterilization and is an oxidation product of furfural which is an important chemical feedstock.
In case for the HUMPPA-COPEC campaign only one sequence was found covering 18 days out of 31 days. Thus the behaviour of the VOCs within this sequence was compared to the rest of the days. It should be noted that the behaviour of temperature and relative humidity are similar between sequence 40 and the rest of the data. The wind speed shows a different pattern as can be seen in Figure 7.19.



(a) Hourly boxplot for temperature.



(b) Hourly boxplot for relative humidity.

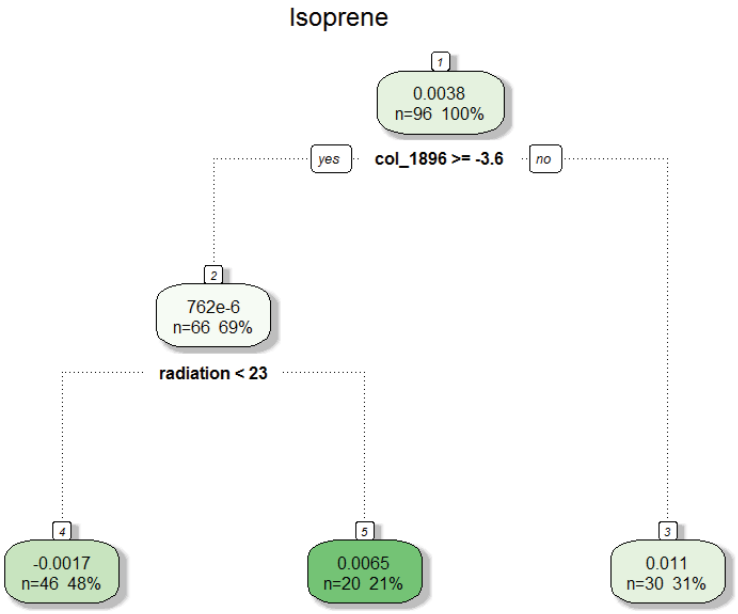


(c) Hourly boxplot for wind speed.

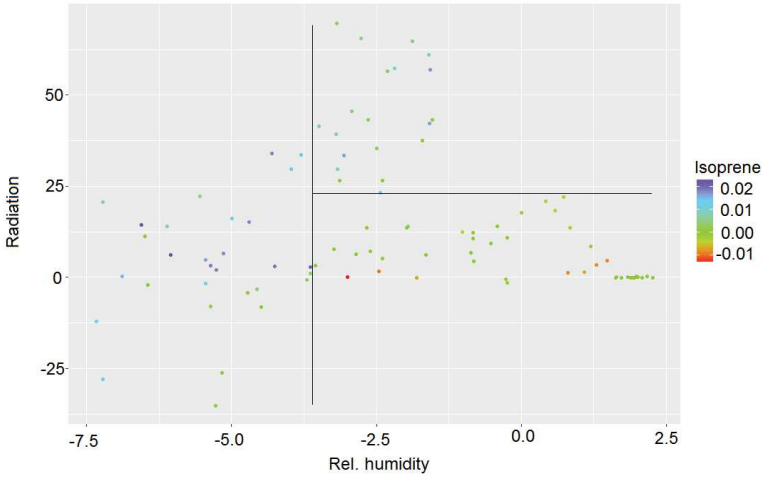
Figure 7.19.: Hourly boxplots for temperature (top right), relative humidity (top left) and wind speed (bottom right) for sequence 40 versus the rest.

Using regression tree models to estimate the influence of the meteorological parameters on the VOCs resulted in similar conclusion as for the CYPHEX data. In Figure 7.20 the representation of the tree model for isoprene is shown. It can be seen that the first variable used for splitting is the difference in relative humidity and the second one the difference in radiation. On the right side in Figure 7.20 the partition of the of the tree model is shown with the humidity and radiation variable. If the difference in humidity gets more negative (higher humidity for the rest of the days) the difference in isoprene rises. Thus the lower the relative humidity the higher the isoprene mixing ratio. This variable splits the data into 31% and 69% of the total data. For the larger part the data is divided using the radiation variable. If the radiation rises the isoprene mixing ratio gets higher. This makes sense since the emission of isoprene is triggered by light.[73] One should bear in mind that this data has a time resolution of 10 minutes and that for sequence 40 which covers about 16 hours of the day 96 data point are available. Therefore, the results are more susceptible to noise.

The rest variables all use the difference in temperature and relative humidity as splitting criteria. Only methanol uses the difference in wind speed for splitting. None of the measured VOCs use the wind direction or the air mass origin as a splitting variable.



(a) Tree model representation.



(b) Scatter plot with humidity and radiation with colour coded values for isoprene.

Figure 7.20.: Representation of the regression tree model for isoprene (top panel) for the HUMPPA-COPEC data. The lower panel shows the scatter plot with relative humidity and radiation with a colour scale for isoprene (ranging from purple for a value of ca. 0.02 to red for a value of ca. -0.01).

7.5. Conclusion

The results show that the algorithm suggested by Guimarães and Mörchen is able to identify patterns in atmospheric time series that are hidden from simple visual inspection. This was demonstrated using two different data sets. It is robust towards noise and only a few parameters must be adjusted. This unsupervised method is useful for identifying patterns in continuous and categorical data. Here we mainly focus on the extraction of patterns from meteorological variables but trace gas time series are possible and was shown for DMS identifying the onset of the sea breeze. A useful extension might be including the factors of a principal component analysis (PCA). The PCA would transform the time series of the measured VOCs into data containing the main variance resulting in time series resembling the behaviour of biogenic or anthropogenic compounds. This would allow in conjunction with meteorological variables to extract patterns which would apply for all biogenic or anthropogenic compounds.
Through the use of dendrograms it is easy to inspect differences in the behaviour of VOCs between two sequences visually. The use of these dendrograms is limited by the amount of variables since the cluster affiliation and the change of VOCs between different clusters becomes unclear. In conjunction with the evaluation of the meteorological variables and their influence on the VOCs interesting insights into the origin and fate of each VOC can be gained.
A promising alternative to hierarchical clustering for assigning VOC to similar groups is the use the classification models. Here we used a randomForest classifier but other classifiers are possible. The PTR-TOF-MS is able to detect hundreds of compounds. Most of the measured masses are unknown and no specific molecule can be assigned to the measured mass-to-charge ratio. The analysis of the similarity between an unknown mass and known compound may lead to new understandings.
Further work includes the implementation of a time window sliding over the time series. This time window could be adjusted to different length and would also allow to capture sequences of a shorter time length. This issue is problematic with the presented approach since the discretization into three intervals is very coarse and thus mainly daily patterns are found. A finer discretization showed no promising results. The implementation of a query algorithm is possible allowing the extraction patterns happening before a chosen variable occurs at “low” or “high” abundance.

8. Conclusion

Humans emit numerous volatile organic compounds (VOCs) into air via skin and breath. These emissions can depend on various factors such as nutrition, sporting activity and also the emotional state. Here we present results of a novel experiment measuring VOCs and CO_2 in a movie theater.
The movie theater provides the opportunity of simultaneously measuring a large group of people (20×230 attendees per film resulting in a total of 13000 measured people) under the same conditions. Secondly, these crowd measurements represent the average emission from a wide cross section of society and are less susceptible to individual behaviour or to similar behavioural pattern from a group of people.
The measurements were performed with a CO_2-monitor and a PTR-TOF-MS. The PTR-TOF-MS enables the real-time measurement of many volatile organic compounds from endogenous and exogenous sources. It was shown that the main breath gases like CO_2, acetone and isoprene mainly depend on the age of the audience. Children (people younger than twelve) generally possess lower emission rates than adults (people of the age of 12 and older). In contrast to that VOCs from exogenous sources strongly vary over the course of day. Their emission rates are subject to behavioural habits like the use of hygiene products or nutritional conventions. For example, the application of daily care products in the morning leads to a constant decrease of the emission rate of decamethylcylopentasiloxane, a typical constituent in those products, over the course of day. In case of methanol the consumption of fruits or juices during breakfast can be the reason for increased emission rates in the screenings before midday. This shows that the emission rates of VOCs can vary strongly between children and adults or over the course of day.
Another influence besides the individual behaviour of humans contributing to the abundance of VOCs in indoor air is the transport of chemicals in this environment. This can be due to dietary habits or use of hygiene products as discussed in the previous paragraph. The environmental conditions such as outdoor ozone mixing ratio and maybe temperature and rainfall can also lead to varying emission rates in indoor environments. We examined the correlation of oxidized VOCs from skin lipids to the mixing ratio of ozone occurring 3 hours before beginning of the film. A positive correlation between 6-MHO and ozone was found. Additionally, increased outside temperature may lead to enhanced evaporation of exogenous compounds included in hygiene products influencing the transport of these compounds into indoor environments.
Compared to the variances due to nutritional habits on a lower magnitude variances in emission rates were found to increase and decrease reproducibly over multiple screenings of the same film, with peaks occurring at the same time point. This observation was decisive for examining potential causal links between the emission of VOCs from the

audience and the audio-visual stimuli occurring in the film. This emission happens on shorter time scales (typically lasting over a few minutes) compared to the discovered age dependence and also the emission from exogenous sources. Over the measurement period, 16 different films of different genre labels (e.g. “action”, “horror”, “comedy”) were recorded. The peaks occurring in the time series of a compound during the screening of the film were induced by the physiological response of the audience. For many films these peaks were characteristic for the film such that the different films were separable by examining visually the behaviour of the VOC trace. To investigate potential connections between VOCs and audio-visual stimuli the films were labelled into sections of suspense, comedy, romantic and many more. Interestingly, the content scenes which could be predicted best from the measured masses were “suspense” and “comedy”. These could be interpreted as intrinsically basic emotions within human beings. Conversely the main breath gases like CO_2 and isoprene and also acetic acid could be predicted best from the annotated content scenes. This significant link seemed reasonable since for example isoprene is known to be emitted in larger amount when the subject holds breath or twitches muscles.

Based on the previous findings indicating a potential connection between the emission of VOCs and a specific content scene the question arose whether this chemical reaction of the audience can be used for the prediction of age ratings of films. Currently, the age rating of a film is decided by a national committee that evaluating several aspects of the film. These aspects include antisocial behaviour, incidences of violence, sex, drug use and bad language. Can this information also be hidden in the time series of the measured VOCs? Considering isoprene, the emission rate increases when an exciting scene is shown due to one or more physiological responses of the audience. Thus the information which can be drawn out from this reaction to the scene can be the height of the peak. Taking into account the whole time series of a film the obtained information can be the number of peaks in the film and the height of these peaks. Interestingly, investigating all measured masses it was found that isoprene exhibits the highest potential in predicting the age rating of a film. Furthermore, the influence of different genre labels of the films (like “action” or “comedy” films) and the different age structure of the audience do not worsen the prediction performance critically. The prediction of the age ratings was made on a different set with unseen films. This indicates that the isoprene trace is able to capture the pattern induced by emotions in the film reflecting the subjective assessment of the committee. Other compounds like CO_2 and acetone are able to distinguish fewer age rating classes than isoprene. However, for single age rating classes they can show higher prediction accuracy than isoprene may reflecting the ability to respond better to specific scenes in this age rating class. It would be interesting if the combination of some of these masses lead to a better prediction accuracy. Additionally, the use of VOCs as indicators in other domains like the detection of psychological stress can be examined. The perception of stress also relies on several environmental conditions and extrinsic stimuli.

The calculated emission rates present robust estimates and can be used for characterizing indoor air influenced by human presence, building ventilation design and comparison of source strength emissions. The investigation of the causal relationship between hu-

man emission of VOCs and audio-visual stimuli might confound the identification of disease biomarkers. Additionally, the chemical response of humans to extrinsic stimuli has practical implications in fields where some objective assessment of groups is required for example in advertising or film industry. It is interesting to think that in future, age classifications of films could be objectively determined by measurement of the reaction of representative test groups. Such approaches can help classification boards on borderline cases or over time as public sensitivities to certain topics change.
The second part of this thesis comprised the use of data mining methods for the analysis of atmospheric time series. The use of pattern identification methods for extracting similar meteorological conditions or repeating behaviour of certain VOCs is useful for the understanding of the chemistry of the VOCs. This was shown for two different data sets taken place in Cyprus and Finland. For both measurement campaigns similar meteorological conditions with a decrease in relative humidity and an increase in wind speed and temperature were found. Additionally, the time series of VOCs can be included into the pattern identification method. For example, the addition of dimethylsulfide to the set of meteorological parameters enabled the objective extraction of time periods when the sea breeze sets in. It was shown for these two measurement campaigns that this pattern identification algorithm is robust towards noise and allows the objective partitioning into smaller time periods of similar behaviour.
The application of clustering methods allowed the division of VOCs into groups of similar behaviour. Many biogenic VOCs such as isoprene and monoterpenes as well as anthropogenic VOC like benzene and toluene were grouped together. Furthermore, representations of hierarchical clustering of different time periods were compared. For the measurement campaign in Cyprus it was shown that biogenic VOCs form similar clusters for different time periods and different meteorological conditions. In contrast, molecules like acetaldehyde change their cluster affiliation for different extracted time periods so we deduce a diel change in the source of acetaldehyde. In order to qualitatively estimate the influence of meteorological parameters and air mass origin on VOCs a decision tree classifier was deployed. This resulted in different groups of VOCs. Biogenic compounds were identified to be mainly dependent on temperature whereas anthropogenic compounds on the air mass origin. For acetaldehyde it was assumed that it has some strong biogenic source as well as some anthropogenic contribution (primary emission and secondary production).

A. Supplement: Real world volatile organic compound emission rates from seated adults and children for use in indoor air studies

Table A.1.: Summary of the measured films.

Film	**USK**	**Screenings**	**Viewers**	**Amount of children**
I'm off then	0	33	4053	0%
Star Wars: The Force Awakens	12	34	3300	0%
Help, I've shrunk my teacher	0	18	988	64%
	sum	**85**	**8341**	**8%**

Table A.3.: Summary of the emission rates of the measured VOCs. All emission rates are presented in [µg $h^{-1}p^{-1}$]. In the last column the calibration method is recorded. "Calibration gas" means that we calibrated this mass with a gas standard. For the monoterpenes we used α-pinene and took the sum of m/z 81.0699 (fragment of many monoterpenes, $C_6H_9^+$) and m/z 137.1325 ($C_{10}H_{17}^+$). "Calibration factor" means that we used the calibration factor of acetone (30.9 [ncps/ppb]) to convert the measured signal into a mixing ratio.

Pr. Mass	Molecule	Adults [µg $h^{-1}p^{-1}$]	Std.dev. [µg $h^{-1}p^{-1}$]	Children [µg $h^{-1}p^{-1}$]	Std.dev. [µg $h^{-1}p^{-1}$]	Cal. method
CO2		$3.0 \cdot 10^7$	$0.5 \cdot 10^7$	$1.8 \cdot 10^7$	$0.6 \cdot 10^7$	Cal. gas
m31.0178	Formaldehyde	207	104	426	375	Cal. gas

m33.0335	Methanol	650	736	1136	984	Cal. gas
m42.0423	Acetonitrile	8	4	9	9	Cal. gas
m43.0542	(iso)Propyl fragment	107	81	321	240	Cal. factor
m45.0335	Acetaldehyde	221	76	252	160	Cal. gas
m47.0491	Ethanol	216	154	116	171	Cal. factor
m57.0699	(iso)Butyl fragment	41	21	52	48	Cal. factor
m59.0491	Acetone	419	96	333	202	Cal. gas
m61.0284	Acetic acid	205	78	357	277	Cal. factor
m63.0084		2	0	1	1	Cal. factor
m63.0263	Methyl mercaptane	7	2	6	5	Cal. gas
m65.0604		5	4	1	5	Cal. factor
m67.0542		6	2	4	3	Cal. factor
m69.0699	Isorpene	166	39	95	59	Cal. gas
m71.0491	Methyl vinyl ketone	8	3	11	8	Cal. gas
m71.0855		8	4	8	10	Cal. factor
m73.0648	Methyl ethyl ketone/Methacrolein	50	29	105	62	Cal. gas
m75.0440	Propionic acid/Hydroxy acetone	19	7	27	18	Cal. factor
m83.0855		22	7	32	23	Cal. factor
m85.0648		3	1	3	2	Cal. factor
m85.1012		3	2	4	4	Cal. factor
m87.0441		8	3	15	11	Cal. factor
m87.0804	Pentanal	3	1	3	2	Cal. factor
m89.0597	Butyric acid	12	4	19	11	Cal. factor
m95.0129		5	5	11	32	Cal. factor
m95.0491	Phenol	10	4	13	10	Cal. factor
m95.0855		15	10	14	12	Cal. factor
m97.0298		10	4	19	14	Cal. factor
m97.1012		7	3	10	7	Cal. factor
m99.0804		5	2	6	4	Cal. factor
m103.0780	Pentanoic acid	4	1	6	5	Cal. factor
m109.1076		11	13	12	12	Cal. factor
m111.1178		5	2	6	5	Cal. factor

m121.0648		28	14	54	45	Cal. factor
m123.1168		4	2	5	3	Cal. factor
m127.1181	6-Methyl-5-heptene-2-one (6MHO)	3	2	5	4	Cal. factor
m131.0850		6	2	4	3	Cal. factor
m133.1012		5	2	5	5	Cal. factor
m135.1168		8	5	3	18	Cal. factor
m137.1325 + m81.0699	Sum of monoterpenes	201	170	189	181	Cal. gas
m143.1067		3	2	3	2	Cal. factor
m143.1430		4	2	8	6	Cal. factor
m145.1150		8	3	2	3	Cal. factor
m153.1274		7	6	6	5	Cal. factor
m155.1430		6	2	4	4	Cal. factor
m157.1577		4	2	5	4	Cal. factor
m159.1363		4	2	4	4	Cal. factor
m177.1608		3	1	3	3	Cal. factor
m201.1857		13	4	9	6	Cal. factor
m205.1951		12	4	12	11	Cal. factor
m207.1768		11	3	9	5	Cal. factor
m217.1734		8	3	4	4	Cal. factor
m221.1568		8	4	13	10	Cal. factor
m235.2056		37	13	12	21	Cal. factor
m355.0698	Fragment of Decamethylpentacycosiloxane	112	104	256	186	Cal. factor

A.1. Description of the PTR-TOF-MS set up

From the 1/4" OD (0.625 cm) main sample line described in the section above, a fraction of 500 mL/min was drawn through a heated PEEK (polyether ether ketone) line (60 °C) with 1/8" OD (0.313 cm) to the PTR-TOF-MS drift tube. The drift tube was operated under 2.20 hPa, a temperature of 60 °C and a drift voltage of 600 V providing an E/N of 137 Td. For mass calibration 1,3,5 Trichlorobenzene was used, permeating through a 1/8" OD Teflon tube in the inlet system. The acquisition time was 30 seconds and mass spectra were recorded ranging from 1 × 400 m/z. The raw PTR-TOF-MS data was evaluated using the PTR-TOF ANALYZER, which is described elsewhere.[103] The signal in counts per second (cps) was normalized by the sum of the intensities of the signals of protonated water (measured on m/z 21 for $H_2{}^{18}O^+$) and

Table A.2.: Summary of the screening hours.

film	screening hour	times
I'm off then	17:30	17
	20:00	16
Star Wars	14:00	20
	18:00	2
	22:30	12
Help, I've shrunk my teacher	11:30	16
	17:20	2

the first water cluster (measured on the m/z 39 for $(H_2{}^{18}O)H_3O^+$). The signal was calculated for a standardized pressure (2.00 hPa) and temperature (20°C). The PTR-TOF-MS was calibrated using a standard gas mixture (Apel-Riemer Environmental Inc., Broomfield, USA) of several VOCs with known mixing ratios. The VOCs included in the calibration gas were methanol, acetonitrile, acetaldehyde, acetone, dimethyl sulfide, isoprene, methyl vinyl ketone, methacrolein, methyl ethyl ketone, benzene, toluene, o-xylene, 1,3,5-trimethylbenzene and α-pinene. However, it could be that in ambient air there are other isomers present but we calibrated the signal only with one compound in our gas standard. In the case of the monoterpenes we used the calibration factor of α-pinene obtained by summing up the signal from m81.0699 (fragment of many monoterpenes, $C_6H_9^+$) and m137.1325 (mass of monoterpenes, $C_{10}H_{17}^+$). To calculate the mixing ratio we used this calibration factor on the sum of these both masses. The signal of decamethylpentasiloxane was converted into the mixing ratio using the most abundant peak on m/z 355.0698 and applying the calibration factor of acetone. The calculated detection limit (3 σ of the noise) ranged from 0.01 ppb (acetaldehyde) to 0.24 ppb (α-pinene) for the measured VOCs. The mixing ratios of VOCs which were not included in the standard gas mixture were calculated with respect to acetone using the sensitivity of the m/z 59.0491. The humidity dependency was studied for various VOCs and it was found that the normalized counts per second varied only slightly ($< 5\%$) for all VOCs in the standard gas mixture for the degree of humidity variation experienced during the opening hours of the cinema. For all calibrated VOCs the sensitivity in normalized counts per second decreased as humidity increased.

B. Supplement: Can the age classification of films be made based on audience breath-chemical emissions?

Table B.1.: Summary of the attendees statistic. The numbers show the average amount of viewers attending the showroom.

FSK 0	FSK 6	FSK 12	FSK 16
Help, I've shrunk my teacher 55	Buddy 104	The Starving Games 52	Counselor 90
I'm off then 122	Dinosaurs 3D 36	Hunger Games: Catching Fire 118	Machete Kills 44
	Walter Mitty 138	Star Wars: The Force Awakens 97	Paranormal Activity: Ghost Dimension 130

B.1. Detailed description of the box model

The modelled mixing ratios for the compounds were calculated by applying a mass-balance-approach. In order to use this model several assumptions must be made including that there is no pathway for mass loss except air exchange and that the emission rate p is small compared to the air exchange rate. It was observed that the air was less effectively mixed in the lower part of the cinema. Thus a mixing factor q must be introduced accounting for the incomplete mixing of the air. The volume of the screening room was 1300 m^3 and the air supply was 6500 m^3/h with identical flows in and out of the showroom provided by the software control of the ventilation system (Instatec Klima-Energietechnik GmbH). Equation 1 shows the ordinary differential equation (ODE) which must be solved optimizing the parameters q the mixing factor and p the emission rate.

$$\frac{\mathrm{dm}}{\mathrm{dt}} = c_{in} \cdot q \cdot r + p - c_{out} \cdot q \cdot r \tag{B.1}$$

In equation B.1, m is the mass of the molecules at time t in the screening room air. The outside air is supplied with a ventilation rate r and a mixing ratio c_{in}. The mixing ratio c_{in} was interpolated from the two surrounding background night time measurements in the absence of people for each VOC. To account for the imperfect air mixing The ventilation rate r is multiplied with the mixing parameter q and the product $(q \cdot r)$ provides a smaller effective room ventilation rate. Consequently, the lower the mixing factor q the worse the mixing of air in the room. The emission rate of a given gas from the audience is given by p. The estimation of the emission rate p and the mixing factor q involves the solving of the ordinary differential equation shown as equation B.1. The optimization was performed using a non-linear least squares method. The estimated constant emission rate p of the VOC and the mixing factor q were used to calculate the mixing ratio of the VOC, which could be seen as the red curve on the left side in Figure 6.1 in the manuscript.

Table B.2.: Summary of the area under ROC curve calculated for all VOCs.

mass	compound	FSK 0	FSK 6	FSK 12	FSK 16
CO2		0.55	0.53	0.75	0.15
m31.0178	Formaldehyde	0.55	0.71	0.48	0.39
m33.0335	Methanol	0.50	0.62	0.36	0.26
m42.0423	Acetonitrile	0.65	0.69	0.33	0.48
m43.0542	(iso)Propyl fragment	0.51	0.67	0.52	0.22
m45.0335	Acetaldehyde	0.45	0.50	0.51	0.14
m47.0152	Ethanol	0.45	0.57	0.56	0.43
m47.0491		0.47	0.57	0.56	0.22
m57.0699	(iso)Butyl fragment	0.61	0.67	0.37	0.35
m59.0491	Acetone	0.55	0.63	0.56	0.13
m61.0284	Acetic acid	0.54	0.54	0.55	0.32
m63.0084		0.52	0.66	0.48	0.58
m63.0263	Methyl mercaptane	0.55	0.76	0.40	0.44
m63.0463		0.47	0.55	0.47	0.60
m65.0215		0.74	0.79	0.40	0.17
m65.0604		0.36	0.64	0.58	0.16
m67.0542		0.40	0.65	0.47	0.52
m69.0335		0.54	0.66	0.56	0.41
m69.0699	Isorpene	0.84	0.74	0.70	0.25
m71.0491	Methyl vinyl ketone	0.47	0.63	0.39	0.37
m71.0855		0.44	0.46	0.53	0.69
m73.0648	Methyl ethyl ketone/Methacrolein	0.60	0.54	0.46	0.74
m75.0440	Propionic acid/Hydroxy acetone	0.46	0.66	0.33	0.36
m77.0536		0.45	0.55	0.57	0.38
m79.0542		0.56	0.62	0.53	0.42

m83.0455		0.38	0.40	0.70	0.60
m83.0855		0.60	0.59	0.49	0.52
m85.0648		0.48	0.45	0.60	0.30
m85.1012		0.47	0.56	0.37	0.29
m87.0441		0.36	0.45	0.49	0.38
m87.0804	Pentanal	0.43	0.46	0.52	0.51
m89.0597	Butyric acid	0.49	0.63	0.59	0.22
m93.0699		0.51	0.56	0.52	0.43
m95.0129		0.51	0.51	0.50	0.47
m95.0491	Phenol	0.53	0.54	0.69	0.29
m95.0855		0.53	0.54	0.70	0.55
m97.0298		0.34	0.47	0.55	0.47
m97.0661		0.38	0.52	0.71	0.55
m97.1012		0.54	0.49	0.52	0.26
m99.0440		0.41	0.46	0.58	0.34
m99.0804		0.44	0.50	0.67	0.44
m101.0597		0.45	0.51	0.60	0.47
m101.0961		0.53	0.52	0.64	0.54
m103.0780	Pentanoic acid	0.54	0.52	0.61	0.63
m107.0855		0.55	0.59	0.59	0.50
m109.1076		0.56	0.70	0.40	0.61
m111.0363		0.40	0.53	0.62	0.61
m111.0800		0.43	0.44	0.45	0.45
m111.1178		0.51	0.64	0.53	0.54
m114.0930		0.54	0.79	0.29	0.32
m115.0754		0.50	0.58	0.54	0.52
m115.1117		0.52	0.44	0.63	0.29
m121.1012		0.51	0.47	0.58	0.36
m135.1168		0.50	0.54	0.56	0.40
m137.1325	Sum of monoterpenes	0.60	0.58	0.64	0.66
m145.1150		0.40	0.33	0.61	0.69
m235.2056		0.51	0.55	0.73	0.48
m355.0698	Decamethylpentacycosiloxane	0.54	0.73	0.59	0.38

Table B.3.: Summary of the standard deviation of the area under curve for all measured VOCs.

mass	compound	FSK 0	FSK 6	FSK 12	FSK 16
CO2		0.08	0.13	0.06	0.10
m31.0178	Formaldehyde	0.10	0.09	0.10	0.10
m33.0335	Methanol	0.06	0.10	0.06	0.12
m42.0423	Acetonitrile	0.11	0.08	0.08	0.15

m43.0542	(iso)propyl fragment	0.07	0.11	0.05	0.15
m45.0335	Acetaldehyde	0.09	0.09	0.07	0.10
m47.0152	Ethanol	0.10	0.13	0.09	0.14
m47.0491		0.09	0.10	0.08	0.13
m57.0699	(iso)Butyl fragment	0.12	0.09	0.11	0.14
m59.0491	Acetone	0.11	0.11	0.09	0.07
m61.0284	Acetic acid	0.09	0.12	0.10	0.17
m63.0084		0.10	0.10	0.10	0.15
m63.0263	Methyl mercaptane	0.15	0.08	0.15	0.20
m63.0463		0.10	0.10	0.10	0.08
m65.0215		0.14	0.12	0.10	0.12
m65.0604		0.06	0.07	0.07	0.06
m67.0542		0.09	0.06	0.11	0.12
m69.0335		0.08	0.14	0.08	0.14
m69.0699	Isorpene	0.07	0.09	0.11	0.13
m71.0491	Methyl vinyl ketone	0.10	0.09	0.10	0.09
m71.0855		0.06	0.09	0.07	0.10
m73.0648	Methyl ethyl ketone/Methacrolein	0.12	0.10	0.12	0.07
m75.0440	Propionic acid/Hydroxy acetone	0.13	0.11	0.10	0.16
m77.0536		0.09	0.09	0.12	0.15
m79.0542		0.08	0.07	0.10	0.11
m83.0455		0.10	0.13	0.12	0.11
m83.0855		0.10	0.07	0.11	0.12
m85.0648		0.10	0.12	0.07	0.10
m85.1012		0.10	0.11	0.11	0.11
m87.0441		0.08	0.12	0.08	0.13
m87.0804	Pentanal	0.10	0.12	0.12	0.14
m89.0597	Butyric acid	0.08	0.13	0.11	0.11
m93.0699		0.05	0.08	0.08	0.12
m95.0129		0.08	0.10	0.06	0.15
m95.0491	Phenol	0.10	0.17	0.12	0.11
m95.0855		0.08	0.12	0.11	0.08
m97.0298		0.07	0.11	0.10	0.17
m97.0661		0.06	0.12	0.11	0.11
m97.1012		0.10	0.11	0.15	0.11
m99.0440		0.09	0.13	0.12	0.15
m99.0804		0.11	0.16	0.14	0.13
m101.0597		0.11	0.10	0.13	0.10
m101.0961		0.10	0.12	0.05	0.12
m103.0780	Pentanoic acid	0.11	0.15	0.13	0.12
m107.0855		0.08	0.10	0.10	0.08
m109.1076		0.11	0.08	0.12	0.05
m111.0363		0.09	0.15	0.14	0.10
m111.0800		0.08	0.11	0.07	0.10

m111.1178		0.11	0.11	0.13	0.16
m114.0930		0.11	0.10	0.09	0.17
m115.0754		0.08	0.13	0.08	0.09
m115.1117		0.08	0.12	0.13	0.11
m121.1012		0.10	0.13	0.07	0.12
m135.1168		0.10	0.12	0.11	0.08
m137.1325	Sum of monoterpenes	0.09	0.10	0.08	0.14
m145.1150		0.09	0.10	0.11	0.12
m235.2056		0.08	0.11	0.10	0.11
m355.0698	Decamethylpentacycosiloxane	0.06	0.08	0.09	0.17

Table B.4.: Summary of the p-value derived from the permutation test for all measured VOCs.

mass	compound	FSK 0	FSK 6	FSK 12	FSK 16
CO2		0.40	0.49	0.09	0.92
m31.0178	Formaldehyde	0.41	0.08	0.57	0.75
m33.0335	Methanol	0.45	0.24	0.85	0.82
m42.0423	Acetonitrile	0.21	0.10	0.88	0.54
m43.0542	(iso)Propyl fragment	0.47	0.16	0.47	0.83
m45.0335	Acetaldehyde	0.68	0.41	0.53	0.93
m47.0152	Ethanol	0.68	0.35	0.37	0.65
m47.0491		0.59	0.34	0.37	0.86
m57.0699	(iso)Butyl fragment	0.31	0.16	0.76	0.68
m59.0491	Acetone	0.30	0.29	0.27	0.95
m61.0284	Acetic acid	0.37	0.51	0.33	0.80
m63.0084		0.44	0.15	0.54	0.41
m63.0263	Methyl mercaptane	0.41	0.08	0.68	0.52
m63.0463		0.58	0.41	0.52	0.32
m65.0215		0.11	0.09	0.67	0.91
m65.0604		0.92	0.16	0.24	0.92
m67.0542		0.78	0.14	0.56	0.50
m69.0335		0.42	0.19	0.32	0.72
m69.0699	Isorpene	0.01	0.05	0.16	0.80
m71.0491	Methyl vinyl ketone	0.59	0.21	0.79	0.70
m71.0855		0.71	0.58	0.44	0.25
m73.0648	Methyl ethyl ketone/Methacrolein	0.32	0.41	0.61	0.15
m75.0440	Propionic acid/Hydroxy acetone	0.56	0.20	0.86	0.72
m77.0536		0.71	0.41	0.37	0.63
m79.0542		0.35	0.22	0.43	0.54
m83.0455		0.79	0.70	0.11	0.37
m83.0855		0.22	0.28	0.50	0.52

m85.0648		0.57	0.61	0.26	0.81
m85.1012		0.68	0.42	0.81	0.81
m87.0441		0.88	0.66	0.55	0.68
m87.0804	Pentanal	0.61	0.55	0.49	0.45
m89.0597	Butyric acid	0.53	0.19	0.30	0.83
m93.0699		0.53	0.34	0.43	0.56
m95.0129		0.44	0.49	0.50	0.53
m95.0491	Phenol	0.42	0.39	0.08	0.86
m95.0855		0.42	0.40	0.08	0.43
m97.0298		0.89	0.58	0.36	0.59
m97.0661		0.86	0.47	0.10	0.46
m97.1012		0.38	0.54	0.46	0.84
m99.0440		0.78	0.58	0.29	0.75
m99.0804		0.60	0.55	0.19	0.57
m101.0597		0.63	0.43	0.22	0.56
m101.0961		0.39	0.44	0.14	0.49
m103.0780	Pentanoic acid	0.40	0.42	0.31	0.25
m107.0855		0.39	0.34	0.31	0.41
m109.1076		0.38	0.11	0.68	0.39
m111.0363		0.76	0.37	0.28	0.28
m111.0800		0.66	0.69	0.66	0.54
m111.1178		0.47	0.19	0.46	0.46
m114.0930		0.39	0.03	0.91	0.78
m115.0754		0.49	0.34	0.41	0.49
m115.1117		0.46	0.63	0.24	0.81
m121.1012		0.52	0.53	0.30	0.67
m135.1168		0.49	0.41	0.38	0.64
m137.1325	Sum of monoterpenes	0.25	0.33	0.16	0.32
m145.1150		0.73	0.85	0.27	0.20
m235.2056		0.47	0.37	0.04	0.54
m355.0698	Decamethylpentacycosiloxane	0.35	0.04	0.27	0.69

The tables S2-S3 (statistic mean and standard deviation for all compounds and all features), tables S7-S8 (ticket sales data for 2013/2014 and 2015/2016) and tables S9-S10 (all measured masses for 2013/2014 and 2015/2016) can be obtained from the author.

Bibliography

[1] IMDb - international movie database.

[2] K. Ackerl, M. Atzmueller, and K. Grammer. The scent of fear. *Neuro Endocrinology Letters*, 23(2):79–84, Apr. 2002.

[3] D. Adamović, J. Doriić, and M. Vojinović-Miloradov. BTEX in the Exhaust Emissions of Motor Vehicles. In *Causes, Impacts and Solutions to Global Warming*, pages 333–342. Springer, New York, NY, 2013.

[4] J. Albrecht, M. Demmel, V. Schöpf, A. M. Kleemann, R. Kopietz, J. May, T. Schreder, R. Zernecke, H. Brückmann, and M. Wiesmann. Smelling Chemosensory Signals of Males in Anxious Versus Nonanxious Condition Increases State Anxiety of Female Subjects. *Chemical Senses*, 36(1):19–27, Jan. 2011.

[5] M. F. Ali and E. D. Morgan. Chemical Communication in Insect Communities: A Guide to Insect Pheromones with Special Emphasis on Social Insects. *Biological Reviews*, 65(3):227–247, Aug. 1990.

[6] A. Amann, M. Corradi, P. Mazzone, and A. Mutti. Lung cancer biomarkers in exhaled breath. *Expert Review of Molecular Diagnostics*, 11(2):207–217, Mar. 2011.

[7] A. Amann, P. Španěl, and D. Smith. Breath analysis: the approach towards clinical applications. *Mini Reviews in Medicinal Chemistry*, 7(2):115–129, Feb. 2007.

[8] M. O. Andreae and P. J. Crutzen. Atmospheric Aerosols: Biogeochemical Sources and Role in Atmospheric Chemistry. *Science*, 276(5315):1052–1058, May 1997.

[9] R. Atkinson. Atmospheric chemistry of VOCs and NOx. *Atmospheric Environment*, 34(12):2063–2101, Jan. 2000.

[10] I. T. Baldwin and J. C. Schultz. Rapid Changes in Tree Leaf Chemistry Induced by Damage: Evidence for Communication Between Plants. *Science*, 221(4607):277–279, July 1983.

[11] J. A. Bernstein, N. Alexis, H. Bacchus, I. L. Bernstein, P. Fritz, E. Horner, N. Li, S. Mason, A. Nel, J. Oullette, K. Reijula, T. Reponen, J. Seltzer, A. Smith, and S. M. Tarlo. The health effects of nonindustrial indoor air pollution. *Journal of Allergy and Clinical Immunology*, 121(3):585–591, Mar. 2008.

[12] R. S. Blake, P. S. Monks, and A. M. Ellis. Proton-Transfer Reaction Mass Spectrometry. *Chemical Reviews*, 109(3):861–896, Mar. 2009.

[13] A. Bracco, F. Falasca, A. Nenes, I. Fountalis, and C. Dovrolis. Advancing climate science with knowledge-discovery through data mining. *npj Climate and Atmospheric Science*, 1(1):4, Jan. 2018.

[14] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997.

[15] M. M. Bradley and P. J. Lang. Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1):49–59, Mar. 1994.

[16] L. Breiman. Random Forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

[17] L. Cappellin, C. Soukoulis, E. Aprea, P. Granitto, N. Dallabetta, F. Costa, R. Viola, T. D. Mark, F. Gasperi, and F. Biasioli. PTR-ToF-MS and data mining methods: a new tool for fruit metabolomics. *Metabolomics*, 8(5):761–770, Oct. 2012.

[18] P. Changkaew and R. Kongkachandra. Automatic movie rating using visual and linguistic information. In *2010 First International Conference on Integrated Intelligent Computing*, pages 12–16, Aug. 2010.

[19] R. J. Charlson, J. E. Lovelock, M. O. Andreae, and S. G. Warren. Oceanic phytoplankton, atmospheric sulphur, cloud albedo and climate. *Nature*, 326(6114):655–661, Apr. 1987.

[20] S.-P. Cheon, S. Kim, S.-Y. Lee, and C.-B. Lee. Bayesian networks based rare event prediction with sensor data. *Knowledge-Based Systems*, 22(5):336–343, July 2009.

[21] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46, Apr. 1960.

[22] S. M. Correa, G. Arbilla, M. R. C. Marques, and K. M. P. G. Oliveira. The impact of BTEX emissions from gas stations into the atmosphere. *Atmospheric Pollution Research*, 3(2):163–169, Apr. 2012.

[23] R. L. Corsi and C.-C. Lin. Emissions of 2,2,4-Trimethyl-1,3-Pentanediol Monoisobutyrate (TMPD-MIB) from Latex Paint: A Critical Review. *Critical Reviews in Environmental Science and Technology*, 39(12):1052–1080, Nov. 2009.

[24] B. d. L. Costello, A. Amann, H. Al-Kateb, C. Flynn, W. Filipiak, T. Khalid, D. Osborne, and N. M. Ratcliffe. A review of the volatiles from the healthy human body. *Journal of Breath Research*, 8(1):014001, 2014.

[25] P. J. Crutzen and a. J. Lelieveld. Human Impacts on Atmospheric Chemistry. *Annual Review of Earth and Planetary Sciences*, 29(1):17–45, 2001.

[26] P. J. Crutzen, J. Williams, U. Pöschl, P. Hoor, H. Fischer, C. Warneke, R. Holzinger, A. Hansel, W. Lindinger, B. Scheeren, and J. Lelieveld. High spatial and temporal resolution measurements of primary organics and their oxidation products over the tropical forests of Surinam. *Atmospheric Environment*, 34(8):1161–1165, Jan. 2000.

[27] J. Davis and M. Goadrich. The Relationship Between Precision-Recall and ROC Curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA, 2006. ACM.

[28] de Gouw J. A., Warneke C., Parrish D. D., Holloway J. S., Trainer M., and Fehsenfeld F. C. Emission sources and ocean uptake of acetonitrile (CH3cn) in the atmosphere. *Journal of Geophysical Research: Atmospheres*, 108(D11), June 2003.

[29] Denise Chen and Jeannette Haviland-Jones. Human Olfactory Communication of Emotion. *Perceptual and Motor Skills*, 91(3):771–781, Dec. 2000.

[30] B. Derstroff, I. Hüser, E. Bourtsoukidis, J. N. Crowley, H. Fischer, S. Gromov, H. Harder, R. H. H. Janssen, J. Kesselmeier, J. Lelieveld, C. Mallik, M. Martinez, A. Novelli, U. Parchatka, G. J. Phillips, R. Sander, C. Sauvage, J. Schuladen, C. Stönner, L. Tomsche, and J. Williams. Volatile organic compounds (VOCs) in photochemically aged air from the eastern and western Mediterranean. *Atmos. Chem. Phys.*, 17(15):9547–9566, Aug. 2017.

[31] M. Dicke, A. A. Agrawal, and J. Bruin. Plants talk, but are they deaf? *Trends in Plant Science*, 8(9):403–405, Sept. 2003.

[32] R. L. Doty. *The great pheromone myth*. The John Hopkins University Press 2010, 2010.

[33] S. Doucet, R. Soussignan, P. Sagot, and B. Schaal. The Secretion of Areolar (Montgomery's) Glands from Lactating Women Elicits Selective, Unconditional Responses in Neonates. *PLOS ONE*, 4(10):e7579, Oct. 2009.

[34] P. Ekman, R. W. Levenson, and W. V. Friesen. Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210, Sept. 1983.

[35] B. Enderby, W. Lenney, M. Brady, C. Emmett, P. Spanel, and D. Smith. Concentrations of some metabolites in the breath of healthy children aged 7-18 years measured using selected ion flow tube mass spectrometry (SIFT-MS). *Journal of Breath Research*, 3(3):036001, 2009.

[36] A. Fabris, F. Biasioli, P. M. Granitto, E. Aprea, L. Cappellin, E. Schuhfried, C. Soukoulis, T. D. Mark, F. Gasperi, and I. Endrizzi. PTR-TOF-MS and data-mining methods for rapid characterisation of agro-industrial samples: influence of

milk storage conditions on the volatile compounds profile of Trentingrana cheese. *J Mass Spectrom*, 45(9):1065–1074, Sept. 2010.

[37] T. Fawcett. ROC Graphs: Notes and Practical Considerations for Researchers. Technical report, 2004.

[38] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[39] J. D. Fenske and S. E. Paulson. Human breath emissions of VOCs. *Journal of the Air & Waste Management Association (1995)*, 49(5):594–598, May 1999.

[40] C. B. Field and et al. *IPCC 2014. Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part A: Global and Sectoral Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press 2014, 2014.

[41] W. Filipiak, V. Ruzsanyi, P. Mochalski, A. Filipiak, A. Bajtarevic, C. Ager, H. Denz, W. Hilbe, H. Jamnig, M. Hackl, A Dzien, and A. Amann. Dependence of exhaled breath composition on exogenous factors, smoking habits and exposure to air pollutants. *Journal of Breath Research*, 6(3):036008, 2012.

[42] G. Freund and P. O'Hollaren. Acetaldehyde concentrations in alveolar air following a standard dose of ethanol in man. *Journal of Lipid Research*, 6(4):471–477, Jan. 1965.

[43] T.-c. Fu. A review on time series data mining. *Engineering Applications of Artificial Intelligence*, 24(1):164–181, Feb. 2011.

[44] A. Gelencsér, K. Siszler, and J. Hlavay. Toluene-Benzene Concentration Ratio as a Tool for Characterizing the Distance from Vehicular Emission Sources. *Environ. Sci. Technol.*, 31(10):2869–2872, Oct. 1997.

[45] S. Gelstein, Y. Yeshurun, L. Rozenkrantz, S. Shushan, I. Frumin, Y. Roth, and N. Sobel. Human Tears Contain a Chemosignal. *Science*, 331(6014):226–230, Jan. 2011.

[46] T. Gierczak, J. B. Burkholder, R. K. Talukdar, A. Mellouki, S. B. Barone, and A. R. Ravishankara. Atmospheric fate of methyl vinyl ketone and methacrolein. *Journal of Photochemistry and Photobiology A: Chemistry*, 110(1):1–10, Oct. 1997.

[47] Y. Gong, W. Wang, S. Jiang, Q. Huang, and W. Gao. Detecting Violent Scenes in Movies by Auditory and Visual Cues. In *Advances in Multimedia Information Processing - PCM 2008*, Lecture Notes in Computer Science, pages 317–326. Springer, Berlin, Heidelberg, Dec. 2008.

[48] M. Graus, M. Müller, and A. Hansel. High resolution PTR-TOF: Quantification and formula confirmation of VOC in real time. *Journal of the American Society for Mass Spectrometry*, 21(6):1037–1044, June 2010.

[49] Griffin Robert J., Cocker David R., Seinfeld John H., and Dabdub Donald. Estimate of global atmospheric organic aerosol from oxidation of biogenic hydrocarbons. *Geophysical Research Letters*, 26(17):2721–2724, Sept. 1999.

[50] V. Gros, C. Gaimoz, F. Herrmann, T. Custer, J. Williams, B. Bonsang, S. Sauvage, N. Locoge, O. dArgouges, R. Sarda-Esteve, and J. Sciare. Volatile organic compounds sources in Paris in spring 2007. Part I: qualitative analysis. *Environmental Chemistry*, 8(1):74–90, Mar. 2011.

[51] G. Guimarães, J.-h. Peter, T. Penzel, and A. Ultsch. A method for automated temporal knowledge acquisition applied to sleep-related breathing disorders. *Artificial Intelligence in Medicine*, page 237, 2001.

[52] G. Guimaraes and A. Ultsch. *A Method for Temporal Knowledge Conversion*. Springer, 1999.

[53] Hamm Stephan and Warneck Peter. The interhemispheric distribution and the budget of acetonitrile in the troposphere. *Journal of Geophysical Research: Atmospheres*, 95(D12):20593–20606, Sept. 2012.

[54] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer, 2009.

[55] Heikes Brian G., Chang Wonil, Pilson Michael E. Q., Swift Elijah, Singh Hanwant B., Guenther Alex, Jacob Daniel J., Field Brendan D., Fall Ray, Riemer Daniel, and Brand Larry. Atmospheric methanol budget and ocean implication. *Global Biogeochemical Cycles*, 16(4):80–1, Dec. 2002.

[56] M. Heil and R. Karban. Explaining evolution of plant communication by airborne signals. *Trends in Ecology & Evolution*, 25(3):137–144, Mar. 2010.

[57] J. Herbig, M. Müller, S. Schallhart, T. Titzmann, M. Graus, and A. Hansel. Online breath analysis with PTR-TOF. *Journal of Breath Research*, 3(2):027004, 2009.

[58] R. A. Hites. Polybrominated diphenyl ethers in the environment and in people a meta analysis of concentrations. *Environmental Science & Technology*, 38(4):945–956, Feb. 2004.

[59] A. T. Hodgson, D. Beal, and J. E. R. McIlvaine. Sources of formaldehyde, other aldehydes and terpenes in a new manufactured house. *Indoor Air*, 12(4):235–242, Dec. 2002.

[60] A. T. Hodgson, J. D. Wooley, and J. M. Daisey. Emissions of volatile organic compounds from new carpets measured in a large-scale environmental chamber. *Air & Waste: Journal of the Air & Waste Management Association*, 43(3):316–324, Mar. 1993.

[61] T. Hoffmann, J. R. Odum, F. Bowman, D. Collins, D. Klockow, R. C. Flagan, and J. H. Seinfeld. Formation of Organic Aerosols from the Oxidation of Biogenic Hydrocarbons. *Journal of Atmospheric Chemistry*, 26(2):189–222, Feb. 1997.

[62] Y. Horii and K. Kannan. Survey of organosilicone compounds, including cyclic and linear siloxanes, in personal-care and household products. *Archives of Environmental Contamination and Toxicology*, 55(4):701–710, Nov. 2008.

[63] S. Hyvönen, H. Junninen, L. Laakso, M. Dal Maso, T. Grönholm, B. Bonn, P. Keronen, P. Aalto, V. Hiltunen, T. Pohja, S. Launiainen, P. Hari, H. Mannila, and M. Kulmala. A look at aerosol formation using data mining techniques. *Atmos. Chem. Phys.*, 5(12):3345–3356, Dec. 2005.

[64] V. A. Isidorov, I. G. Zenkevich, and B. V. Ioffe. Volatile organic compounds in the atmosphere of forests. *Atmospheric Environment (1967)*, 19(1):1–8, Jan. 1985.

[65] K. J. Jardine, R. K. Monson, L. Abrell, S. R. Saleska, A. Arneth, A. Jardine, F. Y. Ishida, A. M. Y. Serrano, P. Artaxo, T. Karl, S. Fares, A. Goldstein, F. Loreto, and T. Huxman. Within-plant isoprene oxidation confirmed by direct emissions of oxidation products methyl vinyl ketone and methacrolein. *Global Change Biology*, 18(3):973–984, Mar. 2012.

[66] A. Jordan, S. Haidacher, G. Hanel, E. Hartungen, L. Märk, H. Seehauser, R. Schottkowsky, P. Sulzer, and T. D. Märk. A high resolution and high sensitivity proton-transfer-reaction time-of-flight mass spectrometer (PTR-TOF-MS). *International Journal of Mass Spectrometry*, 286(2-3):122–128, 2009.

[67] S. Kabinsingha, S. Chindasorn, and C. Chantrapornchai. A movie rating approach and application based on data mining. *International Journal of Engineering and Innovative Technology*, 2(1):77–83, July 2012.

[68] T. Karl, E. Apel, A. Hodzic, D. D. Riemer, D. R. Blake, and C. Wiedinmyer. Emissions of volatile organic compounds inferred from airborne flux measurements over a megacity. *Atmos. Chem. Phys.*, 9(1):271–285, Jan. 2009.

[69] T. Karl, P. Prazeller, D. Mayr, A. Jordan, J. Rieder, R. Fall, and W. Lindinger. Human breath isoprene and its relation to blood cholesterol levels: new measurements and modeling. *Journal of Applied Physiology*, 91(2):762–770, Aug. 2001.

[70] P. Karlson and M. Lüscher. Pheromones: a new term for a class of biologically active substances. *Nature*, 183(4653):55, Jan. 1959.

[71] L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, Sept. 2009.

[72] J. Kesselmeier and A. Hubert. Exchange of reduced volatile sulfur compounds between leaf litter and the atmosphere. *Atmospheric Environment*, 36(29):4679–4686, Oct. 2002.

[73] J. Kesselmeier and M. Staudt. Biogenic Volatile Organic Compounds (VOC): An Overview on Emission, Physiology and Ecology. *Journal of Atmospheric Chemistry*, 33(1):23–88, May 1999.

[74] R. P. Kiene. Dimethyl sulfide metabolism in salt marsh sediments. *FEMS Microbiology Letters*, 53(2):71–78, Mar. 1988.

[75] J. King. Physiological Modeling for Analysis of Exhaled Breath. In *Volatile Biomarkers: Non-invasive diagnosis in Physiology and Medicine*, pages 27–46. Elsivier 2013, 2013.

[76] J. King, A. Kupferthaler, K. Unterkofler, H. Koc, S. Teschl, G. Teschl, W. Miekisch, J. Schubert, H. Hinterhuber, and A Amann. Isoprene and acetone concentration profiles during exercise on an ergometer. *Journal of Breath Research*, 3(2):027006, 2009.

[77] J. King, P. Mochalski, K. Unterkofler, G. Teschl, M. Klieber, M. Stein, A. Amann, and M. Baumann. Breath isoprene: Muscle dystrophy patients support the concept of a pool of isoprene in the periphery of the human body. *Biochemical and Biophysical Research Communications*, 423(3):526–530, July 2012.

[78] Kirstine Wayne, Galbally Ian, Ye Yuerong, and Hooper Martin. Emissions of volatile organic compounds (primarily oxygenated species) from pasture. *Journal of Geophysical Research: Atmospheres*, 103(D9):10605–10619, May 1998.

[79] N. E. Klepeis, W. C. Nelson, W. R. Ott, J. P. Robinson, A. M. Tsang, P. Switzer, J. V. Behar, S. C. Hern, and W. H. Engelmann. The National Human Activity Pattern Survey (NHAPS): a resource for assessing exposure to environmental pollutants. *Journal of Exposure Science and Environmental Epidemiology*, 11(3):231–252, July 2001.

[80] U. Knoll, G. Nakhaeizadeh, and B. Tausend. Cost-sensitive pruning of decision trees. In *Machine Learning: ECML-94*, Lecture Notes in Computer Science, pages 383–386. Springer, Berlin, Heidelberg, Apr. 1994.

[81] R. Kohavi. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In *International Joint Conference on Artifiacl Intelligence*, pages 1137–1143. Morgan Kaufmann, 1995.

[82] A. Krishna, M. O. Lwin, M. Morrin, J. Deighton, and L. Peracchio. Product Scent and Memory. *Journal of Consumer Research*, 37(1):57–67, 2010.

[83] I. Kushch, B. Arendacka, S. Stolc, P. Mochalski, W. Filipiak, K. Schwarz, L. Schwentner, A. Schmid, A. Dzien, M. Lechleitner, V. Witkovsky, W. Miekisch, J. Schubert, K. Unterkofler, and A. Amann. Breath isoprene - aspects of normal physiology related to age, gender and cholesterol profile as determined in a proton transfer reaction mass spectrometry study. *Clinical Chemistry and Laboratory Medicine*, 46(7):1011–1018, 2008.

[84] L. Laffel. Ketone bodies: a review of physiology, pathophysiology and application of monitoring to diabetes. *Diabetes/Metabolism Research and Reviews*, 15(6):412–426, Dec. 1999.

[85] B. Lamb, H. Westberg, G. Allwine, L. Bamesberger, and A. Guenther. Measurement of biogenic sulfur emissions from soils and vegetation: Application of dynamic enclosure methods with Natusch filter and GC/FPD analysis. *J Atmos Chem*, 5(4):469–491, Dec. 1987.

[86] R. B. Lamorena, S.-G. Jung, G.-N. Bae, and W. Lee. The formation of ultrafine particles during ozone-initiated oxidations with terpenes emitted from natural paint. *Journal of Hazardous Materials*, 141(1):245–251, Mar. 2007.

[87] M. Lechner, B. Moser, D. Niederseer, A. Karlseder, B. Holzknecht, M. Fuchs, S. Colvin, H. Tilg, and J. Rieder. Gender and age specific differences in exhaled isoprene levels. *Respiratory Physiology & Neurobiology*, 154(3):478–483, Dec. 2006.

[88] J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer. The contribution of outdoor air pollution sources to premature mortality on a global scale. *Nature*, 525(7569):367–371, Sept. 2015.

[89] R. W. Levenson. Blood, sweat, and fears: the autonomic architecture of emotion. *Annals of the New York Academy of Sciences*, 1000:348–366, Dec. 2003.

[90] H. Levin. Building materials and indoor air quality. *Occupational Medicine (Philadelphia, Pa.)*, 4(4):667–693, Dec. 1989.

[91] W. Lindinger, J. Taucher, A. Jordan, A. Hansel, and W. Vogel. Endogenous Production of Methanol after the Consumption of Fruit. *Alcoholism: Clinical and Experimental Research*, 21(5):939–943, Aug. 1997.

[92] J. Lines, A. Bagnall, P. Caiger-Smith, and S. Anderson. Classification of Household Devices by Electricity Usage Profiles. In *Intelligent Data Engineering and Automated Learning - IDEAL 2011*, Lecture Notes in Computer Science, pages 403–412. Springer, Berlin, Heidelberg, Sept. 2011.

[93] H. M. Loos, S. Doucet, R. Soussignan, C. Hartmann, K. Durand, R. Dittrich, P. Sagot, A. Buettner, and B. Schaal. Responsiveness of Human Neonates to the Odor of 5α-Androst-16-en-3-one: A Behavioral Paradox? *Chemical Senses*, 39(8):693–703, Oct. 2014.

[94] L. D. Martins, C. F. H. Wikuats, M. N. Capucim, D. S. de Almeida, S. C. da Costa, T. Albuquerque, V. S. Barreto Carvalho, E. D. de Freitas, M. de FÃ¡tima Andrade, and J. A. Martins. Extreme value analysis of air pollution data and their comparison between two large urban regions of South America. *Weather and Climate Extremes*, 18:44–54, Dec. 2017.

[95] S. Mendis, P. A. Sobotka, and D. E. Euler. Pentane and isoprene in expired air from humans: gas-chromatographic analysis of single breath. *Clinical Chemistry*, 40(8):1485–1488, Aug. 1994.

[96] D. B. Millet, A. Guenther, D. A. Siegel, N. B. Nelson, H. B. Singh, J. A. de Gouw, C. Warneke, J. Williams, G. Eerdekens, V. Sinha, T. Karl, F. Flocke, E. Apel, D. D. Riemer, P. I. Palmer, and M. Barkley. Global atmospheric budget of acetaldehyde: 3-D model analysis and constraints from in-situ and satellite observations. *Atmos. Chem. Phys.*, 10(7):3405–3425, Apr. 2010.

[97] P. K. Misztal, C. N. Hewitt, J. Wildt, J. D. Blande, A. S. D. Eller, S. Fares, D. R. Gentner, J. B. Gilman, M. Graus, J. Greenberg, A. B. Guenther, A. Hansel, P. Harley, M. Huang, K. Jardine, T. Karl, L. Kaser, F. N. Keutsch, A. Kiendler-Scharr, E. Kleist, B. M. Lerner, T. Li, J. Mak, A. C. Nolscher, R. Schnitzhofer, V. Sinha, B. Thornton, C. Warneke, F. Wegener, C. Werner, J. Williams, D. R. Worton, N. Yassaa, and A. H. Goldstein. Atmospheric benzenoid emissions from plants rival those from fossil fuels. *Scientific Reports*, 5:12064, July 2015.

[98] T. M. Mitchell. *Machine Learning*. McGraw-Hill Education, New York, 1 edition edition, Mar. 1997.

[99] M. J. Molina and L. T. Molina. Megacities and atmospheric pollution. *Journal of the Air & Waste Management Association (1995)*, 54(6):644–680, June 2004.

[100] F. Mörchen and A. Ultsch. Efficient mining of understandable patterns from multivariate interval time series. *Data Min Knowl Disc*, 15(2):181–215, Oct. 2007.

[101] F. Mörchen, A. Ultsch, and O. Hoos. Extracting Interpretable Muscle Activation Patterns with Time Series Knowledge Mining. *Int. J. Know.-Based Intell. Eng. Syst.*, 9(3):197–208, Aug. 2005.

[102] K. Mukhopadhyay and O. Forssell. An empirical investigation of air pollution from fossil fuel combustion and its impact on health in India during 1973-1974 to 1996-1997. *Ecological Economics*, 55(2):235–250, Nov. 2005.

[103] M. Müller, T. Mikoviny, W. Jud, B. D'Anna, and A. Wisthaler. A new software tool for the analysis of high resolution PTR-TOF mass spectra. *Chemometrics and Intelligent Laboratory Systems*, 127(Supplement C):158–165, Aug. 2013.

[104] C. Nadeau and Y. Bengio. Inference for the Generalization Error. *Machine Learning*, 52(3):239–281, Sept. 2003.

[105] W. W. Nazaroff and C. J. Weschler. Cleaning products and air fresheners: exposure to primary and secondary air pollutants. *Atmospheric Environment*, 38(18):2841–2865, June 2004.

[106] X. Y. Ni, H. Huang, and W. P. Du. Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data. *Atmospheric Environment*, 150:146–161, Feb. 2017.

[107] M. Nicas. Estimating exposure intensity in an imperfectly mixed room. *American Industrial Hygiene Association Journal*, 57(6):542–550, June 1996.

[108] M. Ojala and G. C. Garriga. Permutation Tests for Studying Classifier Performance. *J. Mach. Learn. Res.*, 11:1833–1863, Aug. 2010.

[109] F. F. Paes, H. F. d. C. Velho, and F. M. Ramos. Artificial Neural Networks for Estimating the Atmospheric Pollutant Sources. In *Integral Methods in Science and Engineering*, pages 261–271. Birkhäuser Boston, 2011.

[110] A. Persily and L. de Jonge. Carbon dioxide generation rates for building occupants. *Indoor Air*, Mar. 2017.

[111] J. D. Pleil. Role of exhaled breath biomarkers in environmental health science. *Journal of Toxicology and Environmental Health. Part B, Critical Reviews*, 11(8):613–629, Oct. 2008.

[112] D. Powers. Evaluation: From precision, recall and f-measure to roc, informedness, markedness andcorrelation. *J. Mach. Learn. Res.*, 2:37–63, 2011.

[113] J. R. Quinlan. Induction of decision trees. *Mach Learn*, 1(1):81–106, Mar. 1986.

[114] J. R. Quinlan. Simplifying decision trees. *International Journal of Man-Machine Studies*, 27(3):221–234, Sept. 1987.

[115] T. H. Risby and S. F. Solga. Current status of clinical breath analysis. *Applied Physics B*, 85(2-3):421–426, May 2006.

[116] T. M. Rogers, E. P. Grimsrud, S. C. Herndon, J. T. Jayne, C. E. Kolb, E. Allwine, H. Westberg, B. K. Lamb, M. Zavala, L. T. Molina, M. J. Molina, and W. B. Knighton. On-road measurements of volatile organic compounds in the Mexico City metropolitan area using proton transfer reaction mass spectrometry. *International Journal of Mass Spectrometry*, 252(1):26–37, May 2006.

[117] J. Rosbach, E. Krop, M. Vonk, J. van Ginkel, C. Meliefste, S. de Wind, U. Gehring, and B. Brunekreef. Classroom ventilation and indoor air quality - results from the FRESH intervention study. *Indoor Air*, pages 538–545, Aug. 2015.

[118] T. Saito and M. Rehmsmeier. The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, 10(3):e0118432, Mar. 2015.

[119] B. Schaal and S. Al Ain. Chemical signals selected for newborns in mammals. *Animal Behaviour*, 97(Supplement C):289–299, Nov. 2014.

[120] Schade Gunnar W. and Goldstein Allen H. Seasonal measurements of acetone and methanol: Abundances and implications for atmospheric budgets. *Global Biogeochemical Cycles*, 20(1), Feb. 2006.

[121] F. P. Schiestl. The evolution of floral scent and insect chemical communication. *Ecology Letters*, 13(5):643–656, May 2010.

[122] J. H. Seinfeld and S. N. Pandis. *Atmospheric Chemistry and Physics: From Air Pollution to Climate Change*. John Wiley & Sons, Apr. 2016.

[123] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948.

[124] T. D. Sharkey, A. E. Wiberley, and A. R. Donohue. Isoprene Emission from Plants: Why and How. *Ann Bot*, 101(1):5–18, Jan. 2008.

[125] G. M. Shepherd. The Human Sense of Smell: Are We Better Than We Think? *PLOS Biology*, 2(5):e146, May 2004.

[126] H. C. Shields, D. M. Fleischer, and C. J. Weschler. Comparisons among VOCs Measured in Three Types of U.S. Commercial Buildings with Different Occupant Densities. *Indoor Air*, 6(1):2–17, Mar. 1996.

[127] D. Smith, P. Španěl, B. Enderby, W. Lenney, C. Turner, and S. J. Davies. Isoprene levels in the exhaled breath of 200 healthy pupils within the age range 7-18 years studied using SIFT-MS. *Journal of Breath Research*, 4(1):017101, 2010.

[128] P. Španěl, K. Dryahina, A. Rejšková, T. W. E. Chippendale, and D. Smith. Breath acetone concentration; biological variability and the influence of diet. *Physiological Measurement*, 32(8):N23, 2011.

[129] J. D. Spengler and K. Sexton. Indoor air pollution: a public health perspective. *Science*, 221(4605):9–17, July 1983.

[130] K. Stern and M. K. McClintock. Regulation of ovulation by human pheromones. *Nature*, 392(6672):177, Mar. 1998.

[131] B. G. Stone, T. J. Besse, W. C. Duane, C. D. Evans, and E. G. DeMaster. Effect of regulating cholesterol biosynthesis on breath isoprene excretion in men. *Lipids*, 28(8):705–708, Aug. 1993.

[132] C. Stönner, A. Edtbauer, and J. Williams. Real world volatile organic compound emission rates from seated adults and children for use in indoor air studies. *Indoor Air*, 28(1):1–9, July 2017.

[133] P. Sukul, P. Trefz, J. K. Schubert, and W. Miekisch. Immediate effects of breath holding maneuvers onto composition of exhaled breath. *Journal of Breath Research*, 8(3):037102, 2014.

[134] W. T. Swaney and E. B. Keverne. The evolution of pheromonal communication. *Behavioural Brain Research*, 200(2):239–247, June 2009.

[135] X. Tang, P. K. Misztal, W. W. Nazaroff, and A. H. Goldstein. Siloxanes Are the Most Abundant Volatile Organic Compound Emitted from Engineering Students in a Classroom. *Environmental Science & Technology Letters*, 2(11):303–307, Nov. 2015.

[136] X. Tang, P. K. Misztal, W. W. Nazaroff, and A. H. Goldstein. Volatile Organic Compound Emissions from Humans Indoors. *Environmental Science & Technology*, 50(23):12686–12694, Dec. 2016.

[137] V. Tarvainen, H. Hakola, H. Hellén, J. Back, P. Hari, and M. Kulmala. Temperature and light dependence of the VOC emissions of Scots pine. *Atmos. Chem. Phys.*, 5(4):989–998, Mar. 2005.

[138] R. C. Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2013.

[139] D. Toshniwal. Feature extraction from time series data. *Journal of Computational Methods in Sciences and Engineering*, 9(1,2S1):99–110, Jan. 2009.

[140] C. Turner, P. Španěl, and D. Smith. A longitudinal study of ethanol and acetaldehyde in the exhaled breath of healthy volunteers using selected-ion flow-tube mass spectrometry. *Rapid Communications in Mass Spectrometry*, 20(1):61–68, Jan. 2006.

[141] I. Ueta, Y. Saito, K. Teraoka, T. Miura, and K. Jinno. Determination of Volatile Organic Compounds for a Systematic Evaluation of Third-Hand Smoking. *Analytical Sciences*, 26(5):569–574, 2010.

[142] P. R. Veres, P. Faber, F. Drewnick, J. Lelieveld, and J. Williams. Anthropogenic sources of VOC in a football stadium: Assessing human emissions in the atmosphere. *Atmospheric Environment*, 77:1052–1059, Oct. 2013.

[143] T. Wang, A. Pysanenko, K. Dryahina, P. Španěl, and D. Smith. Analysis of breath, exhaled via the mouth and nose, and the air in the oral cavity. *Journal of Breath Research*, 2(3):037013, 2008.

[144] T. C. Wang. A study of bioeffluents in a college classroom. *ASHRAE Transactions*, 81(1), 1975.

[145] Z. Wang and C. Wang. Is breath acetone a biomarker of diabetes? A historical review on breath acetone measurements. *Journal of Breath Research*, 7(3):037109, Sept. 2013.

[146] Wen Zhou and Denise Chen. Fear-Related Chemosignals Modulate Recognition of Fear in Ambiguous Facial Expressions. *Psychological Science*, 20(2):177–183, Feb. 2009.

[147] C. J. Weschler. Changes in indoor pollutants since the 1950s. *Atmospheric Environment*, 43(1):153–169, Jan. 2009.

[148] C. J. Weschler and W. W. Nazaroff. Dermal Uptake of Organic Vapors Commonly Found in Indoor Air. *Environmental Science & Technology*, 48(2):1230–1237, Jan. 2014.

[149] C. J. Weschler, A. Wisthaler, S. Cowlin, G. Tamás, P. StrÃ¸m-Tejsen, A. T. Hodgson, H. Destaillats, J. Herrington, J. J. Zhang, and W. W. Nazaroff. Ozone-Initiated Chemistry in an Occupied Simulated Aircraft Cabin. *Environ. Sci. Technol.*, 41(17):6177–6184, Sept. 2007.

[150] M. L. White, R. S. Russo, Y. Zhou, J. L. Ambrose, K. Haase, E. K. Frinak, R. K. Varner, O. W. Wingenter, H. Mao, R. Talbot, and B. C. Sive. Are biogenic emissions a significant source of summertime atmospheric toluene in the rural Northeastern United States? *Atmos. Chem. Phys.*, 9(1):81–92, Jan. 2009.

[151] J. Wicker, N. Krauter, B. Derstorff, C. Stönner, E. Bourtsoukidis, T. Klüpfel, J. Williams, and S. Kramer. Cinema Data Mining: The Smell of Fear. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 1295–1304, New York, NY, USA, 2015. ACM.

[152] M. C. Wildermuth and R. Fall. Light-Dependent Isoprene Emission (Characterization of a Thylakoid-Bound Isoprene Synthase in Salix discolor Chloroplasts). *Plant Physiology*, 112(1):171–182, Sept. 1996.

[153] J. Williams, J. Crowley, H. Fischer, H. Harder, M. Martinez, T. Petäjä, J. Rinne, J. Bäck, M. Boy, M. Dal Maso, J. Hakala, M. Kajos, P. Keronen, P. Rantala, J. Aalto, H. Aaltonen, J. Paatero, T. Vesala, H. Hakola, J. Levula, T. Pohja, F. Herrmann, J. Auld, E. Mesarchaki, W. Song, N. Yassaa, A. Nölscher, A. M. Johnson, T. Custer, V. Sinha, J. Thieser, N. Pouvesle, D. Taraborrelli, M. J. Tang, H. Bozem, Z. Hosaynali-Beygi, R. Axinte, R. Oswald, A. Novelli, D. Kubistin, K. Hens, U. Javed, K. Trawny, C. Breitenberger, P. J. Hidalgo, C. J. Ebben, F. M. Geiger, A. L. Corrigan, L. M. Russell, H. G. Ouwersloot, J. Vilà-Guerau de Arellano, L. Ganzeveld, A. Vogel, M. Beck, A. Bayerle, C. J. Kampf, M. Bertelmann, F. Köllner, T. Hoffmann, J. Valverde, D. Gonzalez, M.-L. Riekkola, M. Kulmala, and J. Lelieveld. The summertime Boreal forest field measurement intensive (HUMPPA-COPEC-2010): an overview of meteorological and chemical influences. *Atmos. Chem. Phys.*, 11(20):10599–10618, Oct. 2011.

[154] J. Williams and J. Pleil. Crowd-based breath analysis: assessing behavior, activity, exposures, and emotional response of people in groups. *Journal of Breath Research*, 10(3):032001, 2016.

[155] J. Williams, C. Stönner, J. Wicker, N. Krauter, B. Derstroff, E. Bourtsoukidis, T. Klüpfel, and S. Kramer. Cinema audiences reproducibly vary the chemical composition of air during films, by broadcasting scene specific emissions on breath. *Scientific Reports*, 6:25464, May 2016.

[156] A. Wisthaler, G. Tamas, D. P. Wyon, P. Strom-Tejsen, D. Space, J. Beauchamp, A. Hansel, T. D. Märk, and C. J. Weschler. Products of Ozone-Initiated Chemistry in a Simulated Aircraft Environment. *Environ. Sci. Technol.*, 39(13):4823–4832, July 2005.

[157] A. Wisthaler and C. J. Weschler. Reactions of ozone with human skin lipids: Sources of carbonyls, dicarbonyls, and hydroxycarbonyls in indoor air. *Proceedings of the National Academy of Sciences*, 107(15):6568–6575, Apr. 2010.

[158] P. Wolkoff, P. A. Clausen, B. Jensen, G. D. Nielsen, and C. K. Wilkins. Are We Measuring the Relevant Indoor Pollutants? *Indoor Air*, 7(2):92–106, June 1997.

[159] T. D. Wyatt. *Pheromones and animal behavior: Chemical signals and signatures*. Cambridge University Press 2014, 2014.

[160] T. D. Wyatt. The search for human pheromones: the lost decades and the necessity of returning to first principles. *Proc. R. Soc. B*, 282(1804):20142994, Apr. 2015.

[161] K. P. Wyche, P. S. Monks, K. L. Smallbone, J. F. Hamilton, M. R. Alfarra, A. R. Rickard, G. B. McFiggans, M. E. Jenkin, W. J. Bloss, A. C. Ryan, C. N. Hewitt, and A. R. MacKenzie. Mapping gas-phase organic reactivity and concomitant secondary organic aerosol formation: chemometric dimension reduction techniques for the deconvolution of complex atmospheric data sets. *Atmos. Chem. Phys.*, 15(14):8077–8100, July 2015.

[162] C. J. Wysocki and G. Preti. Facts, fallacies, fears, and frustrations with human pheromones. *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology*, 281A(1):1201–1211, Nov. 2004.

[163] A. M. Yañez-Serrano, A. C. Nölscher, E. Bourtsoukidis, B. Derstroff, N. Zannoni, V. Gros, M. Lanza, J. Brito, S. M. Noe, E. House, C. N. Hewitt, B. Langford, E. Nemitz, T. Behrendt, J. Williams, P. Artaxo, M. O. Andreae, and J. Kesselmeier. Atmospheric mixing ratios of methyl ethyl ketone (2-butanone) in tropical, boreal, temperate and marine environments. *Atmos. Chem. Phys.*, 16(17):10965–10984, Sept. 2016.

[164] A. M. Yañez-Serrano, A. C. Nölscher, J. Williams, S. Wolff, E. Alves, G. A. Martins, E. Bourtsoukidis, J. Brito, K. Jardine, P. Artaxo, and J. Kesselmeier. Diel

and seasonal changes of biogenic volatile organic compounds within and above an Amazonian rainforest. *Atmos. Chem. Phys.*, 15(6):3359–3378, Mar. 2015.

[165] R. J. Yokelson, S. P. Urbanski, E. L. Atlas, D. W. Toohey, E. C. Alvarado, J. D. Crounse, P. O. Wennberg, M. E. Fisher, C. E. Wold, T. L. Campos, K. Adachi, P. R. Buseck, and W. M. Hao. Emissions from forest fires near Mexico City. *Atmos. Chem. Phys.*, 7(21):5569–5584, Nov. 2007.

[166] B. Yuan, M. Shao, J. de Gouw, D. D. Parrish, S. Lu, M. Wang, L. Zeng, Q. Zhang, Y. Song, J. Zhang, and M. Hu. Volatile organic compounds (VOCs) in urban air: How chemistry affects the interpretation of positive matrix factorization (PMF) analysis. *J. Geophys. Res.*, 117(D24):D24302, Dec. 2012.

[167] M. J. Zaki. SPADE: An Efficient Algorithm for Mining Frequent Sequences. *Machine Learning*, 42(1-2):31–60, Jan. 2001.

[168] M. J. Zaki and W. Meira. *Data mining and analysis fundamental concepts and algorithms*. Cambridge University Press, 2014.

[169] Y. Zhai, Z. Rasheed, and M. Shah. Semantic classification of movie scenes using finite state machines. *IEE Proceedings - Vision, Image and Signal Processing*, 152(6):896–901, Dec. 2005.

[170] Y. Zhao, S. Wang, L. Duan, Y. Lei, P. Cao, and J. Hao. Primary air pollutant emissions of coal-fired power plants in China: Current status and future prediction. *Atmospheric Environment*, 42(36):8442–8452, Nov. 2008.

[171] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg. Movie Genre Classification via Scene Categorization. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, pages 747–750, New York, NY, USA, 2010. ACM.

List of Figures

List of Tables