European Journal of Wood and Wood Products
https://doi.org/10.1007/s00107-024-02092-0

ORIGINAL ARTICLE

Prediction of volatile organic compound emission from Norway spruce: a chemometric approach combining FT-NIR and TD-GC-MS

Ingrid Bakke[1] · Kelly Peeters[2] · Roland Kallenborn[3] · Anders Q. Nyrud[1]

Received: 18 December 2023 / Accepted: 2 May 2024

Abstract

Continuous and non-invasive analytical methods, such as Fourier transform near-infrared (FT-NIR), are increasingly utilized across various industries, generating substantial data with valuable insights. This study explored the prediction of volatile organic compound (VOC) emission from Norway spruce (*Picea abies*) building materials using a chemometric approach that combined FT-NIR spectroscopy and gas chromatography-mass spectrometry (GC-MS) analysis. VOC emission from various spruce materials (cross-laminated timber, surface-treated interior spruce panel, and untreated interior spruce panel) was measured using GC-MS, alongside the collection of FT-NIR data from the wood surface. By employing multivariate statistical analysis and predictive modeling techniques, the study found a clear potential of NIR-based models in predicting emission of three key VOCs, *α*-pinene, hexanal, and benzaldehyde, from spruce building materials. However, the suggested approach showed prediction uncertainty, largely due to a small data set. Refining and validating this chemometric approach necessitate larger data sets and analysis incorporating a broader range of VOCs. For the proposed approach to replace GC-MS in routine applications, further analysis is needed due to the requirement of comprehensive VOC quantification.

1 Introduction

Wood is a renewable material and has certain advantages that distinguish it from other common construction materials. The ability of trees to sequester carbon through photosynthesis, the many application areas of wood and the fact that wood products can be reused and recycled makes it a sustainable and renewable resource, if properly managed (Pajchrowski et al. 2014). However, wood and wood building products release VOCs, compounds which primarily arise from resin substances produced during secondary metabolism in trees, but are also emitted from anthropogenic sources such as substances for surface treatment of wood building products (Fineschi et al. 2013; Englund 1999). VOCs are defined as organic substances with low boiling points (50–100 °C to 240–260 °C), present in gaseous form at ambient conditions (EPA 2023). Accumulation of VOCs in poorly ventilated indoor spaces raises concern about their impact on human health (Adamová et al. 2020; Asif et al. 2022; Alapieti et al. 2020). As a result, the assessment of indoor air quality has become increasingly important, considering that human exposure to VOCs depends on their concentration in indoor environments (Kotzias 2021). Two trends in the building sector are of particular interest with regard to indoor air quality and material VOC emission. The building sector faces the challenge of reducing their carbon footprint in order to reach global sustainability goals, and one of the strategies for reaching these goals is to substitute building products that have high carbon footprints with timber-based products (Ali et al. 2020). In addition to this, a large percentage of the carbon footprint of a building is related to the occupational stage, and strategies for reducing energy use in this phase include reducing fresh air volume in buildings and utilizing natural ventilation (Peng 2016). The trend toward more airtight buildings and energy-efficient heating and ventilation systems can significantly influence VOC concentration levels (Persson et al. 2019; Ge et al. 2023). Consequently, it is increasingly important to examine the impact of wood building materials and their VOC emission on indoor air quality to ensure sustainable

✉ Ingrid Bakke
ingrid.m.bakke@nmbu.no

[1] Faculty of Environmental Sciences and Natural Resource Management, Norwegian University of Life Sciences, 1433 Ås, Norway

[2] InnoRenew CoE, 6310 Izola, Slovenia

[3] Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, 1433 Ås, Norway

Published online: 14 May 2024

built environments, both with regard to carbon footprint and to indoor air quality.

Coniferous wood species emit a number of different VOCs, with volatile terpenes being the most prominent group, followed by aldehydes (Englund 1999; Pohleven et al. 2019). While indoor environments solely influenced by wood-related VOCs typically pose minimal health risks, their presence can affect perceived indoor air quality, particularly when they exceed odor thresholds (Alapieti et al. 2020). Differentiating between anthropogenic VOCs originating from human activity and biogenic VOCs originating from plant metabolism, particularly for in situ environmental samples, remains challenging (Pohleven et al. 2019; Asif et al. 2022; Kwon et al. 2007; Guo 2011). A standard for assessment of VOC emission in indoor environments, NS-EN 16516:2017+A1:2020, published by the European Committee for Standardization (CEN), involves the measurement of VOC emissions and is used in many European labelling schemes of building materials (Standards Norway 2020). The use of eco-labeled building materials has demonstrated a reduction in VOC concentration in buildings, as shown in a Swedish study of indoor air quality conducted in preschools (Persson et al. 2019). As the chamber test duration in the standard is 28 days, data collection of VOC emission from building materials is time-consuming and costly. Faster and more easily implemented analytical approaches, such as NIR technology, could be explored as an alternative to the chamber emission test.

Since the 1990s the pulp industry has been using NIR to optimize the pulping process (Meder 2016). In some cases in-line NIR instrumentation is also used to investigate the untreated surface of wood. However, a recent study using in-line instruments in sawmills identified limitations of time and varying spectral responses because of transverse flow of timber and different orientation of the wood surfaces (radial and tangential) (Meder 2016). NIR spectroscopy offers advantages such as high speed and minimal sample preparation, making it an excellent tool for rapid screening and qualitative analysis (Lavine and Kwofie 2021). However, its low specificity when dealing with complex matrix samples renders it unsuitable for quantitative analysis when used alone (Beć et al. 2020). Nevertheless, when combined with other analytical techniques, the strengths of different methods can be utilized, and the spectral data of organic materials can provide a general fingerprint for various characteristics. Previous studies have explored the use of spectral data from wood surfaces in combination with multivariate statistical methods for the identification and characterization of wood samples (Hwang et al. 2016; Lande et al. 2010; Flæte and Haartveit 2004; Abe et al. 2022; Cozzolino 2014; Schimleck and Tsuchikawa 2021). For example, Flæte and Haartveit (2004) demonstrated that NIR spectroscopic data, in conjunction with multivariate analysis, adequately predicted decay resistance in Scots pine (*Pinus sylvestris*) exposed to the brown rot fungus *Poria placenta*. This study successfully utilized data mining techniques to predict specific characteristics of wood samples, achieving high correlations between measured and predicted data using partial least squares regression (PLSR). Abe et al. (2022) could distinguish heartwood from sapwood within a single species, suggesting a relationship between NIR spectra of conifer species and their characteristic chemical composition of secondary metabolites. This hypothesis indicates a relationship between the NIR spectra of conifer species and their signature chemical composition of resin compounds such as VOCs. In addition to the aforementioned considerations, machine-learning techniques have shown promise in predicting VOC emissions from wood building products and estimating human-related VOC emissions (Liu et al. 2023).

The objective of this study was to investigate predictability of VOC emission from Norway spruce (*Picea abies*) building materials using NIR spectroscopic data and multivariate regression methods.

2 Materials and methods

In this study, absorbance data from FT-NIR and VOC emission data from TD-GC-MS were collected for Norway spruce (*Picea abies*) surfaces, to evaluate the predictive power of absorbance data concerning VOC emission. Wood samples of 20 mm × 20 mm × 14 mm were analyzed in order to fit the optics window of the NIR instrument and micro chamber emission tests were performed to fit the small wood samples and simultaneously reduce experiment time compared to a large scale chamber test.

2.1 Wood sample collection

Three types of Norway spruce (*Picea abies*) building materials were collected from two different production facilities located in eastern Norway (as shown in Fig. 1) between March 27th and March 29th 2023. All samples were suitable for indoor paneling, with a moisture content of approximately 12%. The sample types collected were cross-laminated timber (CLT), surface treated spruce interior panel (SSP) and untreated spruce interior panel (USP). SSP was treated with a light colored and waterborne lacquer. The spruce wood in all three products originated from plots in the eastern part of Norway, from sawmill facilities within a radius of 100 km (see Fig. 1). Samples were collected from production facilities shortly after production as this was advised in the European standard for determination of emissions to indoor air, NS-EN 16516:2017+A1:2020 (Standards Norway 2020).

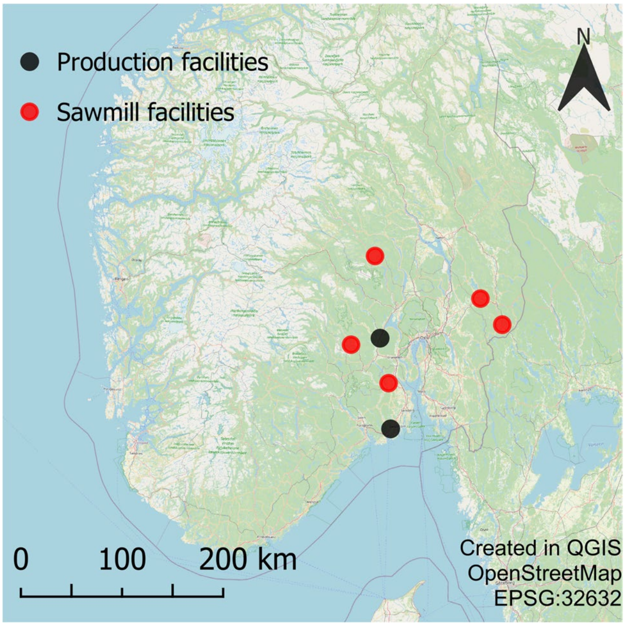

Fig. 1 Location of sawmill facilities (red) and production facilities (black) where Norway spruce building materials were produced and collected

The latest surface cutting of the materials was January 2023 for the SSP and March 2023 for USP and CLT. Test specimens for NIR and GC-MS analysis were prepared by sawing wood samples into cubes of 20 mm × 20 mm × 14 mm, maintaining the surface which would be exposed to the indoor environment. The appearance of the three spruce building materials are shown in Fig. 2. Test specimens were covered by aluminium foil and plastic wrapping to reduce volatilization during transportation and between experiments. Efforts were made to include extremes of faults such as small knots and adhesive joints in the wood surface, but samples with large knots or resin pockets were excluded. This was because of the small sample surface and as an effort to study realistic samples with regard to interior wall panel. Both radial and tangential oriented samples were included in the study to represent realistic absorbance and VOC emission data with regard to in-line production.

The number of wood samples analyzed on FT-NIR and GC-MS can be seen in Table 1. Density and moisture content of the wood samples was also measured, and is presented in Table 1 as mean values for each product type (the complete data set is available in supplement 1 Table S3). Moisture content was measured after drying at 103 °C until constant weight (maximum ±0.5% of previous weight).

2.2 VOC emission measurement

Quantitative analysis of the VOC emission was performed by GC-MS (7890B gas chromatograph 5977B mass spectrometer, Agilent, Santa Clara, CA, USA) with thermal desorption injection (TD 3.5+, Gerstel, Mülheim an der Ruhr, Germany). Wood specimens were placed in micro chambers (μ-CTE250, Markes, Offenbach am Main, Germany) with a loading factor of 3.5 m^2/m^3 and a flow rate of inert gas (N_2 5.0, Istrabenz plini, Koper, Slovenia) through the chambers of 50 mL/min, which corresponded to a ventilation rate of 26 air exchanges per hour. All edges and backside surfaces of the wood specimens were covered by low-emitting aluminium tape. Emission chambers were kept at 23 °C and approximately 50% relative humidity, by introducing ionized

Fig. 2 Example surface of CLT (**a**), SSP (**b**) and USP (**c**) without faults and with faults such as adhesive finger joint in CLT (**d**) and knots in SSP (**e**) and USP (**f**)

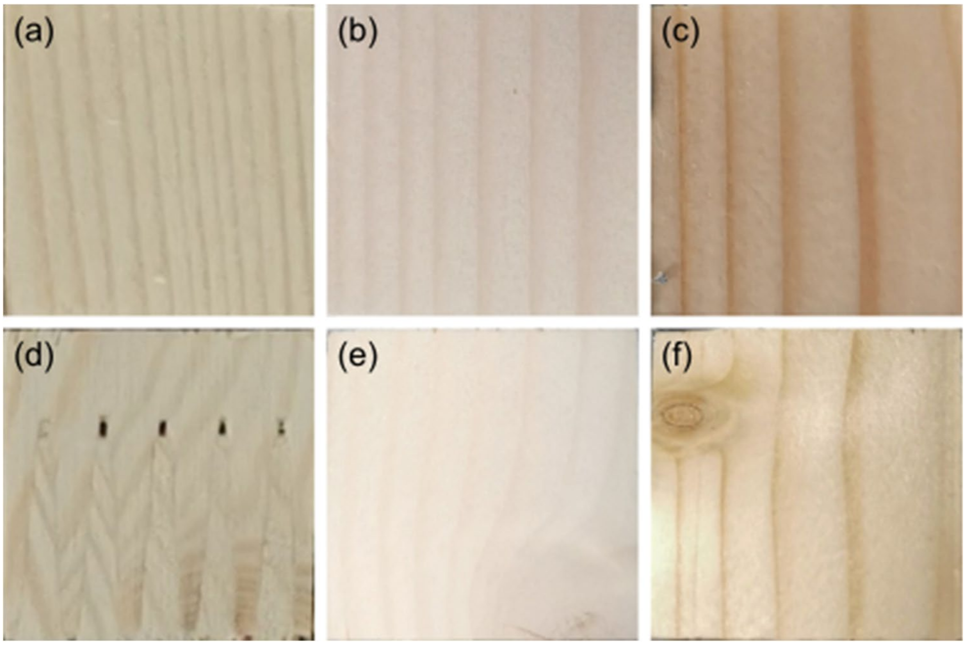

Table 1 Mean and standard deviation (SD) of density and moisture content (MC) of wood samples, as well as the number (N) of wood samples analysed by FT-NIR and TD-GC-MS

	Density ± SD [kg/m^3]	MC ± SD [%]	FT-NIR, N	GC-MS, N
CLT	470 ± 56	10.6 ± 0.82	70	8
SSP	500 ± 31	12.4 ± 0.34	70	8
USP	470 ± 38	12.2 ± 0.63	70	8
Total	480 ± 45	12 ± 1.0	210	24

Table 2 Temperature gradient program used in the GC oven during VOC analysis

Temp. [°C]	Rate [°C/min]	Hold [min]
30	–	1
220	10	–
300	30	4

water ahead of the chambers. After 20 min of equilibration, air samples of 4 L were collected at the chamber outlet, onto conditioned sorbent tubes with 300 mg Tenax TA. Blank samples were collected from empty chambers.

Sorbent tubes were subsequently desorbed in the thermal desorption GC inlet by heating from 30 to 250 °C at a rate of 150 °C/min and holding for 10 min. The desorbed sample was further trapped on the cooled injection inlet (CIS4, Gerstel, Mülheim an der Ruhr, Germany), which was kept in split mode with a split ratio of 10:1 for 0.5 min. The sample was further injected onto the GC column by heating from 10 to 280 °C at 10 °C/s and holding for 1 min.

The GC was run with a flow rate of 1.2 mL/min and an HP5-MS column (60 m, 0.25 mm, 0.25 µm, Agilent, Santa Clara, CA, USA). The temperature program used in the GC oven is shown in Table 2. The MS transfer line was kept at 300 °C and the ion source temperature at 230 °C, in electron ionization (EI) mode with an electron beam energy of 70 eV. The mass range scanned was 35–300 m/z, and the gain factor for signal amplification was set to 1. The acquisition type was performed in scan mode, allowing for the detection of a range of mass-to-charge ratios (m/z).

Eleven analytical standards were used for calibration (toluene, hexanal, furfural, ethylbenzene, *m*-xylene, *o*-xylene, α -pinene, benzaldehyde, β-pinene, β-myrcene and 3-carene), as well as D8-toluene for internal standard calibration. All analytical standards were diluted in dichloromethane, and after inspection, the solvent delay was set to 6 min. GC-MS data was processed using the Masshunter Qualitative Agile 2 method performed on total ion chromatograms. Compound identification was performed by search in the NIST17 database with 10 hits and minimum score of 60. An alkane standard was used to calculate the Kovats retention indices of both target and non-targeted compounds. See supplement 1 (Tables S1 and S2) for more information about equipment and analytical standards.

2.3 NIR data collection

Spectroscopic measurements were performed on an FT-NIR spectrometer (MPA II, Bruker Optics, Ettlingen, Germany), and data was collected in May 2023. The wood surface of 20 mm × 20 mm was placed on the sample tray and scanned for absorbance with a Quartz TE-InGaAs integrating sphere. Each sample was rotated 90 degrees clockwise three times, and the average of the four resulting scans was used for further analysis. This corresponded to a total number of 840 NIR scans, which were further averaged to a total number of 210 averaged scans, one for each wood sample. The spectral region used was 14,000 cm^{-1}–3950 cm^{-1} with a resolution of 8 cm^{-1}, and a phase resolution of 32 scans.

2.4 Multivariate predictive modeling

Pre-processing of NIR data, quantification of VOCs and visualization was performed in R (version 4.3.1, CRAN), while statistical analysis and prediction modeling was performed in JMP Pro (version 16.0.0, SAS). Multivariate statistical analysis was employed due to the large number of variables in spectroscopic data compared to the number of samples (Hastie et al. 2009). Scattering and baseline distortion was performed to account for effects of different particle sizes, crystalline structures and other anatomical features (such as grain orientation and heartwood/sapwood ratio). This was done by applying multiplicative scatter correction (MSC) to the average absorbance spectra within each sample type (Miller and Igne 2021). After investigation of the optimal smoothing procedure (testing 17, 25 and 31 smooting points and first and second derivatives) the resulting spectra were smoothed using a Savitsky–Golay (SG) filter with 31 smoothing points and a second-degree polynomial. This was done to improve the resolution of overlapping bands and enhance important peaks in the spectra. Prediction models were developed using unprocessed, MSC treated and SG filtered data sets to evaluate the best fit with regard to pre-processing methods.

Pre-processing of the NIR data was based on the entire data set, while prediction models were developed using 24 samples with VOC data. Partial least squared regression (PLSR) models to predict VOC emission were constructed using absorbance and individual VOC emission data (Hastie et al. 2009). The data was divided into training (70%) and validation (30%) sets, and different validation methods were explored (manual division with inclusion of each material type in validation set (see details in supplement 1 Table S4), 4-fold and random holdback with 25%). Manual division was performed in order to guarantee inclusion of VOC data for at least two samples of each building material in the training and validation sets to obtain the different concentration levels necessary for linear regression. In addition to PLSR models, bootstrap forest was tested for generation of a collection of decision trees based on random sampling from the data set. This was executed to compensate for the small data set of 12–24 VOC measurements (Hastie et al. 2009; Saccenti and Timmerman 2016). Bootstrap models were divided into the 70% training and 30% validation sets manually chosen. All 3525 independent variables (NIR wavenumbers) were included as terms in the bootstrap forest models and a minimum of 5 sample splits were performed. A standard of 100 trees in the forest was used and the minimum and maximum splits per tree were 5 and 2000, respectively. The sampling rate of bootstrap samples from real samples was set to 5 or 10. The linear correlation coefficient (r^2) and root mean squared error (RMSE) were used to assess the performance of all models. In order to assess the importance of specific wavenumber bands in the prediction of individual VOCs variable importance projections (VIP) from the best PLSR models were extracted and compared to previous findings.

3 Results and discussion

3.1 VOC emission

Among the eleven target analytes, the four compounds α-pinene, hexanal, benzaldehyde and 3-carene were quantified within the linear range in at least two of the three spruce materials. Minimum, median and maximum emission concentrations of these four VOCs are shown in Table 3. See supplement 1 (Tables S5 and S6) for all VOC concentrations, along with median significance measures and calibration results.

The VOCs exhibited relatively modest emission levels from all samples, ranging from 0.3 to 14 µg/m^3. The remaining target compounds were either not detected or detected below their individual linear calibration range. The surface treated spruce panel (SSP) emitted higher median levels of hexanal, benzaldehyde and 3-carene than the other spruce materials, along with a greater TVOC emission. The untreated spruce panel (USP) yielded the highest concentrations of the most characteristic conifer VOC, α-pinene. CLT generally emitted low concentrations of all the quantified target compounds. Given that CLT had the most recently cut surface, it was expected to emit higher VOC levels due to low surface age; however, factors like natural wood variation and adhesive presence might underlie this phenomenon. The geographical proximity of the spruce trees used for the building materials, although within a certain range, does not fully eliminate growth variations. In addition to this, the proportion of heartwood to sapwood might differ, especially between CLT and the two panel products SSP and USP, as the products were planed differently and in different factories. Emission of hexanal and α-pinene is previously found to be higher from sapwood of spruce than from heartwood (Czajka et al. 2020). However, for the purpose of implementing these prediction models in spruce panel industry the models must be independent of the ratio of heartwood to sapwood in spruce surfaces. Total VOC emission was semi-quantified by toluene equivalents, by including all peak areas between n-hexane and n-hexadecane in each total ion chromatogram. Although TVOC is a term that gives minimal information concerning the speciation of VOCs emitted it was included in the models to investigate whether TVOC could be predicted from NIR absorbanse data.

In the context of predictive modeling, variance in the dependent variable is essential for the model to yield meaningful insights. When the variance in the dependent variable is minimal, constructing a predictive model that offers valuable insights or accurate predictions becomes challenging (Hastie et al. 2009). In such instances, even if a model can account for a significant portion of the variance, its practical use may be limited. Between group variance of VOC

Table 3 Emission concentrations of individual VOCs and TVOC from the three building materials. ChiSquare values and significance level from Wilcoxon or Kruskal–Wallis tests indicate the variation in emission concentration between material types

[μg/m^3]		α-Pinene	3-Carene	Hexanal	Benzaldehyde	TVOC
CLT	Min	0.4	0.1	2		40
	Median	1	0.2	3		57
	Max	1	0.3	14		73
SSP	Min	0.3	0.1	6	3	100
	Median	1	0.3	9	3	140
	Max	8	9	12	3	200
USP	Min	3	0.1	3	0.3	67
	Median	6	0.5	4	1	97
	Max	12	0.6	5	3	210
ChiSquare		14	4	12	6	16
p value		0.001	0.1	0.003	0.02	0.0004

emission was therefore tested to assess the distribution of VOC emission included in the prediction. As the VOC data was not normally distributed, Wilcoxon or Kruskal-Wallis tests were performed to regard the levels of VOC emission from the building materials. VOC emission concentrations varied between the materials, as seen from the ChiSquared values and significance levels in Table 3. α-Pinene, hexanal and TVOC showed significantly different median emission concentrations in at least two of the three building material types, with a p value of less than 0.003.

3.2 NIR characterization

Average untreated NIR spectra of the three building materials are depicted in Fig. 3, while spectra processed by MSC and SG are shown in Fig. 4. NIR data collected and treated with the different pre-treatment methods are given in supplement 1 (Tables S7–S10). The wavenumber range was adjusted for illustrative purposes in Fig. 4 (8000–3950 cm^{-1}), although the regression models employed the full range of 14,000–3950 cm^{-1}. From principle component analysis (PCA) and visual inspection of the resulting biplot, one outlier was identified and excluded from the unprocessed NIR data. More specifically, one of the four absorbance spectra of sample CLT55 contained only negative values, and was therefore excluded. Preprocessing was performed on the whole data set of 210 wood samples (70 samples from each material type). Although this can introduce some unwanted influence it was necessary in this case on account of the small data set and time-consuming GC-MS analysis. Inclusion of all 210 samples in pre-processing of the NIR data allowed for a broader range of variation included in the data set. However, the presumption that the quantitative emission data of the 24 test specimens subjected to micro chamber experiments was representative of the whole data set of 210 samples was made in this analysis.

Despite NIR's moderate chemical specificity due to overlapping bands, some trends were present in the NIR

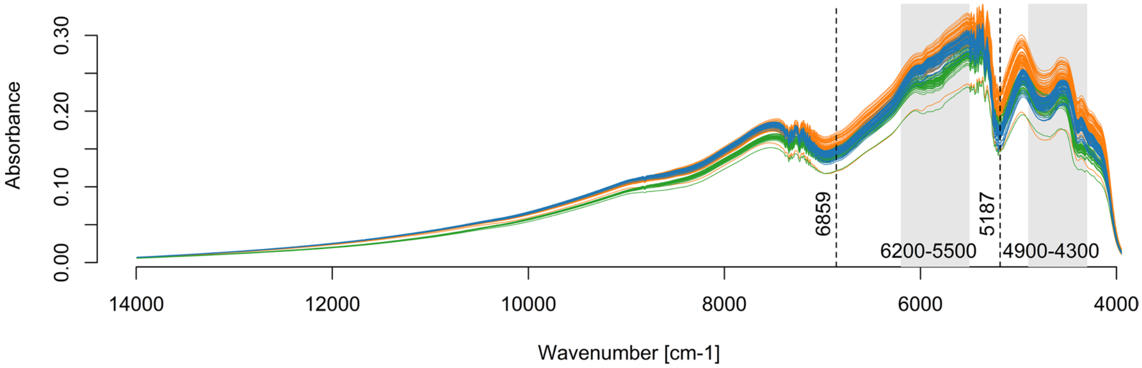

Fig. 3 Average absorbance of each spruce sample surface in the near infra-red region (14,000–3950 cm^{-1}). Orange lines indicate CLT samples; green lines indicate SSP samples; blue lines indicate USP samples; black vertical lines indicate moisture related wavenumbers; shaded areas indicate wavenumber regions with visible variation between material types

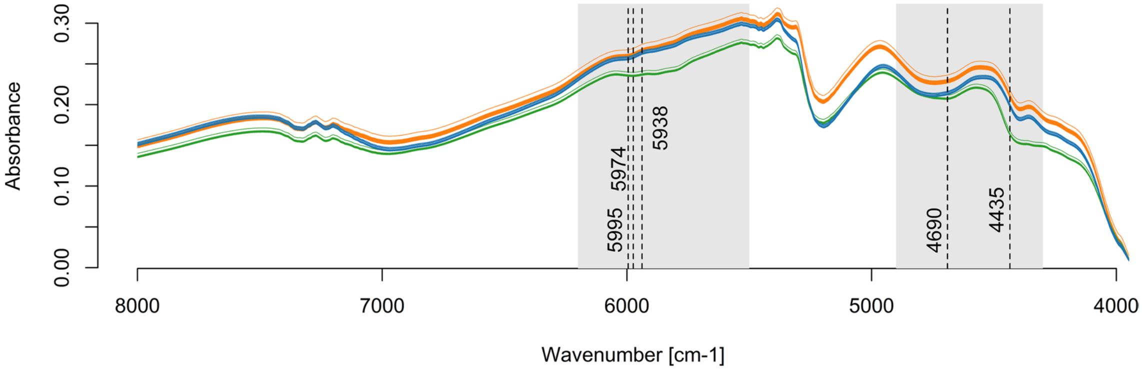

Fig. 4 MSC and SG filtered absorbance of spruce in the near infra-red region (8000–3950 cm^{-1}). Orange lines indicate CLT samples; green lines indicate SSP samples; blue lines indicate USP samples; black vertical lines indicate wavenumbers of interest; shaded areas indicate wavenumber regions with visible variation between material types (6200–5500 cm^{-1}; 4900–4300 cm^{-1})

spectra which can be discussed and compared to previous findings (Schwanninger et al. 2011; Sandak et al. 2016). Absorbance data of the spruce building materials all followed similar trends, although SSP absorbance deviated form CLT and USP in certain regions. Key spectral regions linked to wood moisture's OH-groups were previously observed around 5187 cm^{-1} (O–H deformation) and 6859 cm^{-1} (O–H stretching combination) (Ercioglu et al. 2018; Schwanninger et al. 2011). However, at these bands CLT had slightly higher peak intensity than SSP and USP despite its lower moisture content, and the bands may not be significantly correlated with moisture content in spruce in this experiment.

Spectral distinctions were particularly evident between the three spruce materials in regions 6200–5500 cm^{-1} and 4900–4300 cm^{-1}. The utilization of MSC and SG techniques accentuated these differences, as highlighted in Fig. 4. The region at 6200–5800 cm^{-1} was related to the second overtone of C-H stretching vibrations in carbohydrates and lipids in biological samples (Beć et al. 2020). Variation in this region might be related to hexanal or benzaldehyde emission as aldehydes are secondary oxidation products of lipids (Grebenteuch et al. 2021; Risholm-Sundman et al. 1998). Zhang and Lee (1997) found that hexanal concentration on a silica gel could be predicted by NIR reflectance data in the wavenumber range of 6061-5587 cm^{-1}. As SSP had slightly shaper peaks in this range compared to CLT and USP, this region might contain information about the hexanal emission from spruce.

The presence of aromatic compounds in lignin and extractives, such as benzaldehyde, is also associated with this region of the NIR spectrum. Schwanninger et al. (2011) reported that the band at 5974 cm^{-1} was related to lignin and extractives and arose from the first overtone C-H stretch of aromatic compounds. Additionally, there were some indications that bands at 5995 cm^{-1} and 5938 cm^{-1} were combination bands of the aromatic C–H stretch. However, further examinations of these bands are still required as they might arise from first overtone C–H stretching from methyl groups. According to Schwanninger et al. (2011), these bands (5995 cm^{-1}, 5974 cm^{-1} and 5938 cm^{-1}) also correlated strongly with the band at 4690 cm^{-1}, which arose from the combination of aromatic C–H stretch and C=C stretch in lignin and extractives. The lower wavenumber region of 4900–4300 cm^{-1} might therefore also be correlated to aromatic extractive compounds in wood.

Guo et al. (2006) measured absorbance of α-pinene in the NIR region and reported a high absorbanse intensity in the region 4400–4300 cm^{-1}, due to combination stretching and bending of C–H bonds. From Fig. 4, however, both USP and CLT samples have more intense peaks in this wavenumber region, while only USP had significantly higher α-pinene emissions. This region was also related to C–H stretching combination of $–CH_2$ and $–CH_3$ groups in plants, and the variation may arise from other differences in the spruce structures (Ercioglu et al. 2018).

The two regions containing noise around 7200 cm^{-1} and 5400 cm^{-1} were smoothed during the pre-processing, although without significant disruption to the remaining spectra. These regions could be related to phenolic hydroxyl groups in lignin (7092 cm^{-1} and 6913 cm^{-1}) and to combination of O–H stretch and second overtone C-O stretch in cellulose (5464 cm^{-1}), respectively (Schwanninger et al. 2011). In general, the NIR spectra of SSP varied from those of CLT and USP, although there was no visually obvious correlation between the main differences in individual VOC emission and the discussed spectral regions. Multivariate analysis was therefore necessary to interpret the predictive power of the absorbanse data.

3.3 Prediction of VOC emission

Prediction of VOC emission from the three spruce building materials was performed using VOC emission data as dependant variables ($N = 12–24$) and NIR absorbance data as independant variables ($p = 3525$). Three consecutive pre-treatment methods of NIR data were used to create data sets for the regression (unprocessed, MSC and MSC + SG), where the unprocessed NIR data was averaged spectral results from the four measurements. Two methods for linear regression were utilized, namely PLSR and bootstrap forest. Prediction models were evaluated by number of factors, F, r^2 and RMSE. Models resulting in a correlation coefficient of the validation set below 0.5 were excluded from the results, which are shown in Tables 4 and 5. High correlation coefficients and relatively low error terms were obtained from multiple models for α-pinene, benzaldehyde and hexanal. However, model deficits like possible overfitting, differences in linearity in validation and training sets and small data set indicate uncertainty in these predictions. 3-Carene was not predicted adequately ($r^2 < 0.50$) by either PLSR or bootstrap forest models and TVOC was only predicted quite poorly by PLSR models. The poor predictability of 3-carene was thought to be a result of low emission concentrations and high measurement uncertainty for this analyte.

α-Pinene was most accurately predicted with a PLSR model using the unprocessed NIR data and 4-fold validation. This model had a linearity of 0.86 in the validation set and a relative error term of 1.1 µg/m^3. The error was comparable to the median concentration of α-pinene from both CLT and SSP samples, indicating high uncertainty in the prediction model at these low concentrations. α -Pinene prediction did not improve by random sampling in bootstrap models. α-Pinene was predicted with a linearity of 0.79 and error of 1.4 µg/m^3 in the validation set with bootstrap forest and a sampling rate of 10 from the

Table 4 PLSR model results from predictive modeling of VOC emission, evaluated by number of factors (F), r^2 and RMSE. Validation was performed either by 4-fold validation, manual division of data set into 70% test and 30% validation set or random holdback of 25% of the data set for validation

Dependent variable	Validation	F	r^2	RMSE	N	Data set	Pretreatment
α-Pinene	4-Fold	6	0.86	1.1	7	Validation	Unprocessed
	4-Fold	6	0.78	1.7	15	Training	Unprocessed
	4-Fold	3	0.69	1.7	7	Validation	MSC
	4-Fold	3	0.55	2.4	15	Training	MSC
	4-Fold	3	0.69	1.7	7	Validation	MSC + SG
	4-Fold	3	0.55	2.4	15	Training	MSC + SG
Benzaldehyde	Manual 70/30	6	0.73	0.55	3	Validation	Unprocessed
	Manual 70/30	6	0.93	0.28	9	Training	Unprocessed
	4-Fold	2	0.53	0.73	3	Validation	MSC
	4-Fold	2	0.59	0.69	9	Training	MSC
	4-Fold	2	0.53	0.73	3	Validation	MSC + SG
	4-Fold	2	0.59	0.69	9	Training	MSC + SG
Hexanal	Holdback 25%	4	0.75	1.4	7	Validation	MSC
	Holdback 25%	4	0.35	2.8	17	Training	MSC
	4-Fold	3	0.72	1.5	7	Validation	MSC + SG
	4-Fold	3	0.40	2.7	17	Training	MSC + SG
	Holdback 25%	3	0.70	1.5	7	Validation	MSC + SG
	Holdback 25%	3	0.41	2.7	17	Training	MSC + SG
TVOC	Holdback 25%	3	0.53	31	7	Validation	MSC
	Holdback 25%	3	0.56	33	17	Training	MSC
	4-Fold	4	0.52	31	7	Validation	Unprocessed
	4-Fold	4	0.57	32	17	Training	Unprocessed

Table 5 Bootstrap forest model results from predictive modeling of VOC emission, evaluated by r^2 and RMSE. Bootstrap sampling rate was set to either 5 or 10

Dependent variable	Sampling rate	r^2	RMSE	N	Division	Pretreatment
α-Pinene	10	0.79	1.4	7	Validation	Unprocessed
	10	1.00	0.05	15	Training	Unprocessed
	5	0.59	1.9	7	Validation	Unprocessed
	5	0.97	0.61	15	Training	Unprocessed
Hexanal	5	0.92	0.77	7	Validation	MSC + SG
	5	0.98	0.45	17	Training	MSC + SG
	5	0.88	0.96	7	Validation	MSC
	5	0.98	0.44	17	Training	MSC
	10	0.87	0.97	7	Validation	MSC
	10	1.00	0.05	17	Training	MSC

unprocessed NIR data. The improvement from sampling rate 5–10 implies the need for a lager data set to improve prediction.

Benzaldehyde was most accurately predicted by a PLSR model with manual division of data for validation (70/30%), using the unprocessed NIR data. The validation set had a linearity of 0.73 and an error of 0.55 μg/m^3. As for α-pinene, the prediction of benzaldehyde did not improve by bootstrap sampling. The fact that α-pinene and benzaldehyde models had highest linearity when built on untreated data might indicate a correlation between emission of these VOCs and physical aspects in the spruce surface, such as density. Scatter correction by MSC and smoothing with an SG filter was generally considered to remove physical aspects like scattering and baseline distortion in NIR data (Windig et al. 2008; Chen et al. 2013). However, the number of factors in the best α-pinene and benzaldehyde models was quite high, at 6 factors, indicating a possible overfitting of the models built using the untreated NIR data set. The number of factors used in the 4-fold validation models of MSC treated data and MSC + SG treated data for both VOCs was much lower, at 3 and 2 for α-pinene and benzaldehyde, respectively. These models resulted in lower r^2, at 0.69 and 0.53 for α-pinene and benzaldehyde, respectively. Combined with the small sample number these models indicated lower certainty in the fundamental correlation.

The best predictability of hexanal was obtained by a bootstrap forest model with a sampling rate of 5. This model had linearity of 0.92 and error of 0.77 µg/m^3 for the validation set, and was built from SG filtered data. The second best prediction model was a bootstrap forest model with MSC treated data, and it was not improved when increasing sampling rate from 5 to 10. The PLSR models of hexanal emission also yielded best results when based on SG or MSC treated data. Here a random holdback validation of 25% of the samples was the best model with linearity 0.75 and an error of 1.4 µg/m^3 in the validation set. All PLSR models for hexanal led to poor linearity in the training set and high error terms compared to median emission values. There were differences in linearity of validation and training sets with both 25% holdback and 4-fold validation of hexanal PLSR models. This indicated that the emission range was not adequately covered for these random data set divisions and is another indication that the data set was too small to obtain complete certainty in the prediction models. The relative error in the best bootstrap forest model, however, was lower than the median of measured hexanal emission, and this was thus considered a more robust model than the best *α*-pinene or benzaldehyde models.

TVOC emission concentrations were predicted with a linearity of 0.53 and an error of 31 µg/m^3 in the validation set by a PLSR model with 25% random holdback of samples, using the MSC treated NIR data. TVOC emission data is by definition nonspecific regarding chemical grouping and speciation of VOCs (Salthammer 2022). As NIR absorbance data contains information about combination and overtones from molecular vibrations which are specific for chemical grouping it is to be expected that TVOC would not be predicted adequately. NIR data was therefore considered to be more suitable for prediction of individual VOC emission from spruce wood than of TVOC.

The two best prediction models were obtained for *α* -pinene and hexanal, and the actual by predicted plots for these two models are shown in Fig. 5. The actual by predicted plot of the best benzaldehyde model was not included as there were too few data points to obtain a meaningful plot.

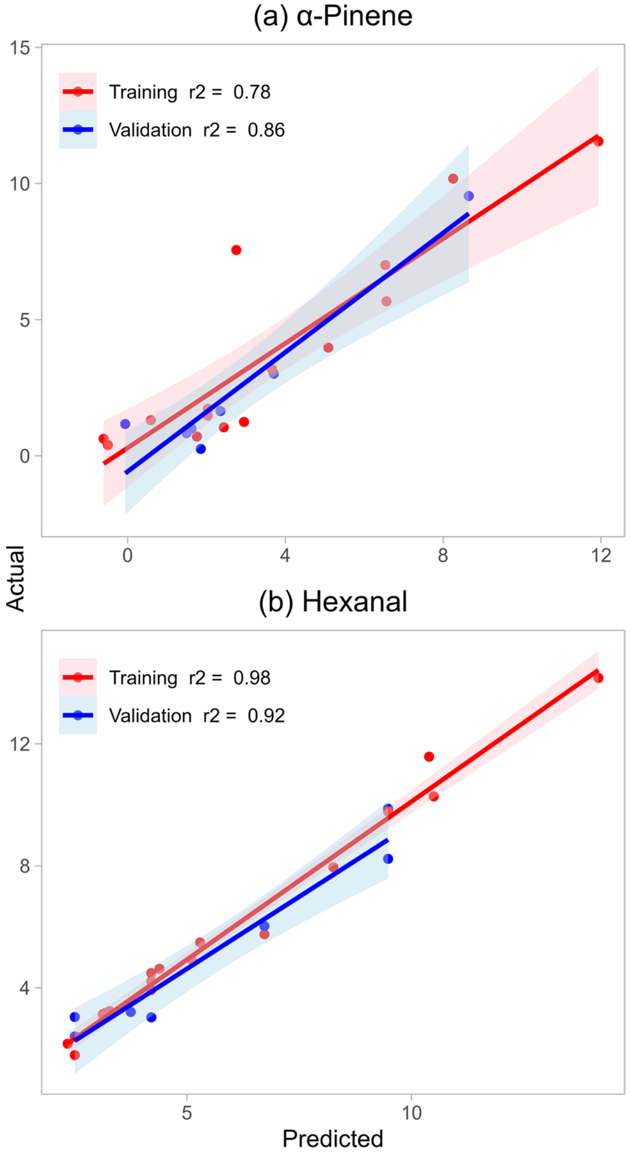

Fig. 5 Actual by predicted plots for **a** *α*-pinene PLSR model using 4-fold validation and unprocessed NIR data and **b** hexanal bootstrap model with sampling rate 5 from SG filtered data. Red shows training data; blue shows validation data

3.4 Variable importance

Variable importance projections (VIP) from the best PLSR models of *α*-pinene, benzaldehyde and hexanal were extracted for further inspection. The independent variables with highest VIP were evaluated with regard to similarities in the NIR spectra and previously reported spectral bands. Variables with high influence on the models were defined as having VIP > 1, and are shown visually in Fig. 6. The range and number of wavenumbers with VIP > 1, along with the wavenumber with highest VIP are listed in Table 6.

Limiting the variables to only wavenumbers with VIP > 1 did not enhance the prediction for any of the three VOCs. PLSR prediction models were constructed using the wavenumber subset shown in Fig. 6 but all r^2 values remained lower than those presented in Table 4. The reduction of wavenumber regions in NIR spectroscopy lacks meaningful impact on analysis time, given that each scan took approximately 20 s.

The regions 6200–5500 cm^{-1} and 4900–4300 cm^{-1} were somewhat correlated with prediction of the three VOCs *α* -pinene, benzaldehyde and hexanal, as variable importance was above 1 for these wavenumber regions. As expected,

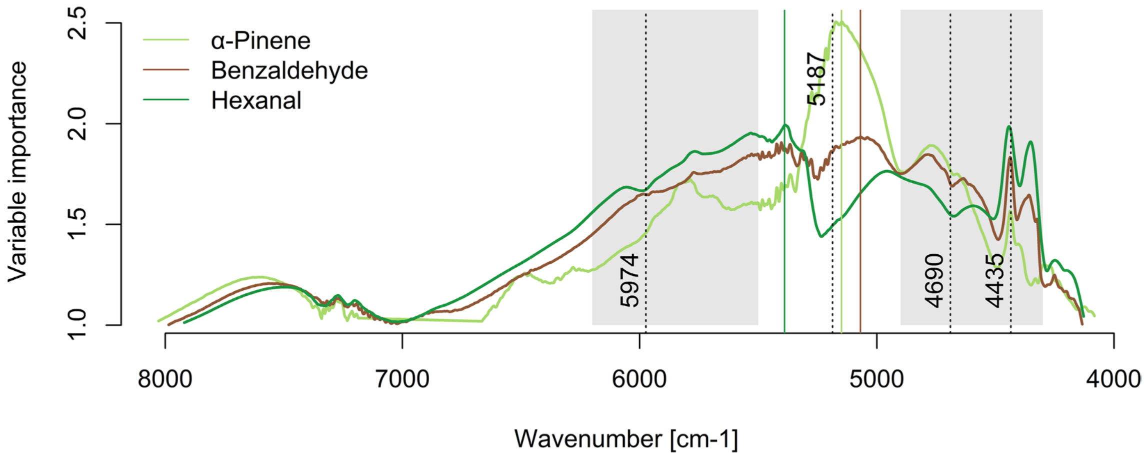

Fig. 6 Variable importance projections above 1 from best PLSR models for α-pinene, benzaldehyde and hexanal. Colored lines indicate highest VIP wavenumbers for each VOC; dotted lines indicate wavenumbers of interest; shaded areas indicate initial wavenumber regions of interest (6200–5500 cm^{-1}; 4900–4300 cm^{-1})

the wavenumbers in region 6200–5500 were more important in the prediction of aldehyde than terpene emission, as this region might be related to emissions of lipid-derived aldehydes (Beć et al. 2020; Grebenteuch et al. 2021). The wavenumbers with maximum prediction importance, however, were all between these two regions at 5069, 5149 and 5389 cm^{-1} for benzaldehyde, α-pinene and hexanal, respectively. The region previously found to be related to hexanal on silica gel, 6061–5587 cm^{-1} (Zhang and Lee 1997), was indeed a large part in predicting hexanal emissions from spruce. The high variable importance at wavenumber band 4435 cm^{-1} might be related to acetyl groups but this band overlaps with a range of wood related bands from cellulose and hemicellulose in softwoods between 4400–4200 cm^{-1} (Schwanninger et al. 2011; Sandak et al. 2016). The same can be said for benzaldehyde, as it had high variable importance projections around 6200–5500 cm^{-1}. This correlated with previous findings, as this region was related to both the second overtone vibration of C–H in lipids and with first overtone aromatic C–H stretches (the band at 5974 cm^{-1}) (Schwanninger et al. 2011). The band at 4690 cm^{-1} was related to aromatic extractives (Schwanninger et al. 2011), and benzaldehyde did show a higher VIP value at this wavenumber than hexanal, although slightly lower than the aliphatic compound α-pinene. The best α-pinene prediction had a clear maximum variable importance around 5150 cm^{-1}, very close to the O–H deformation band of moisture in wood at 5187 cm^{-1} (Ercioglu et al. 2018). The region 4400–4300 cm^{-1} was previously related to combination C–H stretching and bending in α-pinene (Guo et al. 2006), however this region was slightly less relevant in the prediction of α-pinene emissions in the PLSR model.

Table 6 Variable importance projection > 1 for the best PLSR model of α-pinene, benzaldehyde and hexanal

	Range [cm^{-1}]		N	Highest VIP [cm^{-1}]
	Min	Max		
α-Pinene	4083	8028	1197	5149
Benzaldehyde	4128	7920	1331	5069
Hexanal	4134	7986	1352	5389

4 Conclusion

An interdisciplinary approach using FT-NIR, TD-GC-MS and multivariate data analysis is a potential tool for predicting emission of VOCs such as α-pinene, hexanal and benzaldehyde from Norway spruce (*Picea abies*) building materials. The labor-intensive measurement of VOC emissions using TD-GC-MS could potentially be substituted by in-situ NIR measurements coupled with robust prediction models, particularly in cases where NIR technology has already been successfully implemented in wood industry processes. Method evaluation of the VOC prediction approach was not attempted due to low number of samples, and validation of the proposed method is a natural next step in this work. Some of the presented models showed overfitting and low linearity terms, both issues that need to be investigated with a larger data set. Although the combination of micro chamber analysis and NIR analysis is less time-consuming than chamber emission tests following the horizontal European standard (NS-EN 16516:2017+A1:2020), the approach must be tested using chamber sizes within the scope of the standard. Additionally, the list of VOCs of interest in emission measurements increases continuously as new VOCs are detected and evaluated with regard to human toxicology, which poses the need for a wide range of prediction models. While NIR data proves advantageous due to the complexity and specificity of absorbance data, it becomes problematic when combined with nonspecific VOC data such as TVOC. Uncertainty in the measured emission concentrations contributed to relatively high error terms in the prediction models in this study. As the VOC emission concentrations for the majority of the targeted compounds were below quantification ranges, considering a design of study with samples that have higher volatilization, such as Scots pine (*Pinus sylvestris*) interior panel, might be a viable alternative.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00107-024-02092-0.

Acknowledgements The authors would like to thank the European Commission for funding the InnoRenew project (grant agreement #739574), under the H2020 Widespread-2-Teaming program, the Republic of Slovenia (investment funding from the Republic of Slovenia and the European Regional Development Fund), and The Slovenian Research Agency (ARRS) infrastructure program IO-0035. Furthermore, we are thankful to The Norwegian University of Life Sciences (NMBU) for their financial support, which made this research possible. The authors would like to thank Hilde Vinje, Associate Professor and biostatistical advisor at NMBU, for input and evaluation of the statistical approach utilized in this study.

Author contributions Conceptualization: IB, AQN, RK; Methodology: IB and KP; Formal analysis and investigation: IB; Writing—original draft preparation: IB; Writing—review and editing: IB, RK, AQN, KP; Funding acquisition: RK, KP; Supervision: AQN, RK, KP.

Funding Open access funding provided by Norwegian University of Life Sciences

Data availability statement All data supporting the findings of this study is available within the paper and the Supplementary Information. Raw data files are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no financial or non-financial interests that are directly or indirectly related to the work submitted for publication.

References

Abe H, Kurata Y, Watanabe K, Kitin P, Kojima M, Yazaki K (2022) Longitudinal transmittance of visible and near-infrared light in the wood of 21 conifer species. IAWA J 43:403–412. https://doi.org/10.1163/22941932-bja10103

Adamová T, Hradecký J, Pánek M (2020) Volatile organic compounds (VOCs) from wood and wood-based panels: methods for evaluation, potential health risks, and mitigation. Polymer 12:2289. https://doi.org/10.3390/polym12102289

Alapieti T, Mikkola R, Pasanen P, Salonen H (2020) The influence of wooden interior materials on indoor environment: a review. Eur J Wood Prod 78:617–634. https://doi.org/10.1007/s00107-020-01532-x

Ali KA, Ahmad MI, Yusup Y (2020) Issues, impacts, and mitigations of carbon dioxide emissions in the building sector. Sustainability 12:7427. https://doi.org/10.3390/SU12187427

Asif Z, Chen Z, Haghighat F, Nasiri F, Dong J (2022) Estimation of anthropogenic VOCs emission based on volatile chemical products: a Canadian perspective. Environ Manag 71:685–703. https://doi.org/10.1007/s00267-022-01732-6

Beć KB, Grabska J, Huck CW (2020) Near-infrared spectroscopy in bio-applications. Molecules 25:2948. https://doi.org/10.3390/molecules25122948

Chen H, Song Q, Tang G, Feng Q, Lin L (2013) The combined optimization of Savitzky-Golay smoothing and multiplicative scatter correction for FT-NIR PLS models. ISRN Spectrosc 2013:642190. https://doi.org/10.1155/2013/642190

Cozzolino D (2014) Use of infrared spectroscopy for in-field measurement and phenotyping of plant properties: instrumentation, data analysis, and examples. Appl Spectrosc Rev 49:564–584. https://doi.org/10.1080/05704928.2013.878720

Czajka M, Fabisiak B, Fabisiak E (2020) Emission of volatile organic compounds from heartwood and sapwood of selected coniferous species. Forests 11:92. https://doi.org/10.3390/f11010092

Englund F (1999) Emissions of volatile organic compounds (VOC) from wood. Technical report Institutet för träteknisk forskning. https://ri.diva-portal.org/smash/record.jsf?pid=diva2%3A1079819&dswid=9879. Accessed 12 Dec 2023

EPA (2023) Technical overview of volatile organic compounds. Technical report EPA US. https://www.epa.gov/indoor-air-quality-iaq/technical-overview-volatile-organic-compounds. Accessed 12 Dec 2023

Ercioglu E, Velioglu HM, Boyaci IH (2018) Determination of terpenoid contents of aromatic plants using NIRS. Talanta 178:716–721. https://doi.org/10.1016/j.talanta.2017.10.017

Fineschi S, Loreto F, Staudt M, Penuelas J (2013) Diversification of volatile isoprenoid emissions from trees: evolutionary and ecological perspectives. In: Niinemets Ü, Monson RK (eds) Biology, controls and models of tree volatile organic compound emissions, 1st edn, pp 1–20. Springer Dordrecht, Heidelberg

Flæte PO, Haartveit EY (2004) Non-destructive prediction of decay resistance of *Pinus sylvestris* heartwood by near infrared spectroscopy. Scand J For Res 19:55–63. https://doi.org/10.1080/02827580410017852

Ge M, Zheng Y, Zhu Y, Ge J, Zhang Q (2023) Effects of air exchange rate on VOCs and odor emission from PVC veneered plywood used in indoor built environment. Coatings 13:1608. https://doi.org/10.3390/coatings13091608

Grebenteuch S, Kroh LW, Drusch S, Rohn S (2021) Formation of secondary and tertiary volatile compounds resulting from the lipid oxidation of rapeseed oil. Foods 10:2417. https://doi.org/10.3390/foods10102417

Guo H (2011) Source apportionment of volatile organic compounds in Hong Kong homes. Build Environ 46:2280–2286. https://doi.org/10.1016/j.buildenv.2011.05.008

Guo C, Shah RD, Dukor RK, Freedman TB, Cao X, Nafie LA (2006) Fourier transform vibrational circular dichroism from 800 to 10,000 cm^{-1}: near-IR-VCD spectral standards for terpenes and related molecules. Vib Spectrosc 42:254–272. https://doi.org/10.1016/j.vibspec.2006.05.013

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: Data mining, inference, and prediction, 2nd edn. Springer, New York

Hwang SW, Horikawa Y, Lee WH, Sugiyama J (2016) Identification of pinus species related to historic architecture in Korea using NIR chemometric approaches. J Wood Sci 62:156–167. https://doi.org/10.1007/s10086-016-1540-0

Kotzias D (2021) Built environment and indoor air quality: the case of volatile organic compounds. AIMS Environ Sci 8:135–147. https://doi.org/10.3934/environsci.2021010

Kwon KD, Jo WK, Lim HJ, Jeong WS (2007) Characterization of emissions composition for selected household products available in Korea. J Hazard Mater 148:192–198. https://doi.org/10.1016/j.jhazmat.2007.02.025

Lande S, Riel SV, Høibø OA, Schneider MH (2010) Development of chemometric models based on near infrared spectroscopy and thermogravimetric analysis for predicting the treatment level of furfurylated Scots pine. Wood Sci Technol 44:189–203. https://doi.org/10.1007/s00226-009-0278-x

Lavine BK, Kwofie F (2021) Pattern recognition applied to the classification of near-infrared spectra. In: Ciurczak E (ed) Handbook of near-infrared analysis, 4th edition. CRC Press, Oxford, pp 201–208

Liu J, Zhang R, Xiong J (2023) Machine learning approach for estimating the human-related VOC emissions in a university classroom. Build Simul 16:915–925. https://doi.org/10.1007/s12273-022-0976-y

Meder R (2016) Near infrared spectroscopy: seeing the wood in the trees. NIR News 27:26–28. https://doi.org/10.1255/nirn.1580

Miller CE, Igne B (2021) Preprocessing methods in NIR spectroscopy. In: Ciurczak E (ed) Handbook of near-infrared analysis, 4th edition. CRC Press, Oxford, pp 290–312

Pajchrowski G, Noskowiak A, Lewandowska A, Strykowski W (2014) Wood as a building material in the light of environmental assessment of full life cycle of four buildings. Constr Build Mater 52:428–436. https://doi.org/10.1016/j.conbuildmat.2013.11.066

Peng C (2016) Calculation of a building's life cycle carbon emissions based on ecotect and building information modeling. J Clean Prod 112:453–465. https://doi.org/10.1016/j.jclepro.2015.08.078

Persson J, Wang T, Hagberg J (2019) Indoor air quality of newly built low-energy preschools - are chemical emissions reduced in houses with eco-labelled building materials? Indoor Built Environ 28:506–519. https://doi.org/10.1177/1420326X18792600

Pohleven J, Burnard MD, Kutnar A (2019) Volatile organic compounds emitted from untreated and thermally modified wood - a review. Wood Fiber Sci 51:231–254. https://doi.org/10.22382/wfs-2019-023

Risholm-Sundman M, Lundgren M, Vestin E, Herder P (1998) Emissions of acetic acid and other volatile organic compounds from different species of solid wood. Holz Roh- Werkst 56:125–129. https://doi.org/10.1007/s001070050282

Saccenti E, Timmerman ME (2016) Approaches to sample size determination for multivariate data: applications to PCA and PLS-DA of omics data. J Proteome Res 15:2379–2393. https://doi.org/10.1021/acs.jproteome.5b01029

Salthammer T (2022) TVOC - revisited. Environ Int 167:107440. https://doi.org/10.1016/j.envint.2022.107440

Sandak J, Sandak A, Meder R (2016) Assessing trees, wood and derived products with near infrared spectroscopy: hints and tips. J Near Infrared Spectrosc 24:485–505. https://doi.org/10.1255/jnirs.1255

Schimleck LR, Tsuchikawa S (2021) Application of NIR spectroscopy to wood and wood-derived products. In: Ciurczak E (ed) Handbook of near-infrared analysis, 4th edn. CRC Press, Oxford, pp 759–772

Schwanninger M, Rodrigues JC, Fackler K (2011) A review of band assignments in near infrared spectra of wood and wood components. J Near Infrared Spectrosc 19:287–308. https://doi.org/10.1255/jnirs.955

Standards Norway (2020) NS-EN 16516:2017+A1:2020 construction products: assessment of release of dangerous substances: determination of emissions into indoor air. Technical report CEN/TC 351, Standards Norway. Accessed 12 July 2020

Windig W, Shaver J, Bro R (2008) Loopy MSC: a simple way to improve multiplicative scatter correction. Appl Spectrosc 62:1153–1159. https://doi.org/10.1366/000370208786049097

Zhang H-Z, Lee T-C (1997) A novel silica gel adsorption/near-infrared spectroscopic method for the determination of hexanal as an example of volatile compounds. J Agric Food Chem 45:3083–3087. https://doi.org/10.1021/jf970440i