Materials and Structures (2025) 58:255
https://doi.org/10.1617/s11527-025-02755-1

RILEM TC REPORT

Evaluating the reproducibility and consistency of different sample preparation techniques used for ATR-FTIR spectroscopy from the RILEM 295-FBB TG1 round robin test

Johannes Mirwald · Sadaf Khalighi · Aikaterini Varveri · Bernhard Hofko · Dheeraj Adwani · Augusto Cannone-Falchetto · Michael Elwardany · Rita Kleiziené · Katarzyna Konieczna · Maciej Maliszewski · Virginie Mouillet · Sayeda Nahar · Nathalie Piérard · Georgios Pipintakos · Laurent Porot · Kristina Primerano · Aditi Sharma · Pejoohan Tavassoti · Sandra Weigel · Jens Wetekam · Jiqing Zhu

Received: 26 February 2025 / Revised: 22 July 2025 / Accepted: 2 August 2025 / Published online: 19 September 2025

Abstract Attenuated Total Reflection Fourier Transform Infrared spectroscopy has become a popular spectroscopic technique in bituminous binder analysis. However, comparable results are not obtainable yet due to differences in devices, measurement routines, sample preparation procedures, and spectral evaluation. Thus, the Task Group 1 of the RILEM TC 295-FBB: "Fingerprinting bituminous binders using physicochemical analysis" focuses on bringing this method towards pre-standardization. This study evaluates the reproducibility and consistency from round robin test, where 21 participating laboratories performed six different preparation techniques on three different binders in an unaged, short-term, and

Johannes Mirwald and Sadaf Khalighi have contributed equally to this work.

This manuscript was prepared by working group 1 within RILEM TC 295-FBB "Fingerprinting bituminous binders using physicochemical analysis". And further reviewed and approved by all members of the RILEM TC 295-FBB.TC Membership: TC Chair: Bernhard Hofko TC Secretary:Members:Dheeraj Adwani, Waleed Al Nasser, Panos Apostolidis, Johan Blom, Johannes Büchner, Augusto Cannone-Falchetto, Xavier Carbonneau, Alan Carter, Emmanuel Chailleux, Davide Dalmazzo, Eshan Dave, Herve DiBenedetto, Michael Elwardany, Yangming Gao, Victor Garcia Rabadan, Flavien Geisler, Andrea Graziani, Meng Guo, Ankit Gupta, Hmazeh Haghshenas, Bernhard Hofko, Patricija Kara De Maeijer, Sadaf Khalighi, Rita Kleiziene, Katarzyna Konieczna, Jan Król, Peng Li, Yi Li, Bert Jan Lommerts, Maciej Maliszewski, Salvatore Mangiafico, David Mensching, Miomir Miljković, Johannes Mirwald, Mohsen Mhadhbi, Lucas Mortier, Virginie Mouillet, Sayeda Nahar, Jorge Pais, Manfred Partl, Emiliano Pasquini, Nathalie Piérard, Georgios Pipintakos, Kristina Primerano, Jian Qiu, Ali Raman, Shisong Ren, Margarida Sa-Da-Costa, Cesare Sangiorgi, Aditi Sharma, Hilde Soenen, Anand Sreeram, Mayank Sukhija, Yuxuan Sun, Pejoohan Tavassoti, Marjan Tusar, Jan Unterbuchschachner, Stefan Vansteenkiste, Aiaterini Varveri, Kamila Vasconcelos, ThibaultVillette, Dawei Wang, Di Wang, Sandra Weigel, Stefan Werkovits, Jens Wetekamp, Michael Wistuba, Tchedele Langollo Yannik, Yuan Zhang, Fan Zhang, Jiqing Zhu.

J. Mirwald (✉) · B. Hofko
Christian Doppler Laboratory for Chemo-Mechanical Analysis of Bituminous Materials, Institute of Transportation, TU Wien, Gusshausstrasse 28/E230-3, 1040 Vienna, Austria
e-mail: Johannes.mirwald@tuwien.ac.at

S. Khalighi · A. Varveri
Pavement Engineering Section, Engineering Structures Department, Civil Engineering and Geosciences, Delft University of Technology, Stevinweg 1, 2628 CN Delft, The Netherlands

D. Adwani
Department of Civil, Architectural and Environmental Engineering, The University of Texas at Austin, 301 E Dean Keeton St, B.122, Austin, TX 78712, USA

long-term aged state. A total of 6461 spectra were recorded and evaluated for their mean, standard deviation and coefficient of variation (CV) in the spectral region between 1800 and 600 cm^{-1}. The results show that the solid sample preparation methods provide excellent reproducibility, with a coefficient of variation below 2%. Only the solvent method showed a higher coefficient of variation at 7.18%. Outliers with a high CV were detected and categorized into two groups: one where only one of the four samples differed and the other where all 16 spectra showed slight scattering in the overall absorption. The consistency of the method is significantly influenced by the accuracy of sample preparation, which is crucial for minimizing differences in slope, baseline, and noise in the spectra. These findings show the excellent reproducibility of these sample preparation methods and will be further examined to establish universal indices for evaluating effects such as ageing, bringing the method closer towards standardization.

Keywords Bitumen · FTIR spectroscopy · Reproducibility · Round robin test · Material handling · Preparation routine

1 Introduction

Chemical analysis represents a crucial aspect that helps with understanding the properties and performance of bituminous materials used in road construction. Due to its ability to quickly investigate certain chemical moieties, known as functional groups, Fourier Transform Infrared (FTIR) spectroscopy has become one of the most popular chemical analysis techniques used in bitumen research. In organic chemistry, functional groups are described as molecules or atomic groups that significantly determine the properties or reaction behavior of a material. In the context of bitumen, these functional groups play a crucial role in processes like ageing, as reported in early literature by Petersen et al. [1].

Many functional groups in bitumen are infrared (IR) active, which means that they change their dipole momentum upon absorption of infrared light. This absorption causes the molecules to rotate or

A. Cannone-Falchetto
Department of Civil Engineering, Aalto University, Rakentajanaukio 4, 02150 Espoo, Finland

M. Elwardany
FAMU-FSU College of Engineering, 2525 Pottsdamer St., Tallahassee, FL 32310, USA

R. Kleiziené
Road Research Institute of Environmental Engineering Faculty of Vilnius Gediminas Technical University, Sauletekio av. 11, LT-10223 Vilnius, Lithuania

K. Konieczna
Faculty of Civil Engineering, Warsaw University of Technology, Aleja Armii Ludowej 16, 00-637 Warsaw, Poland

M. Maliszewski
Road and Bridge Research Institute, Instytutowa 1, 03-302 Warsaw, Poland

V. Mouillet
Cerema, Univ Gustave Eiffel, UMR MCD, 13100 Aix-en-Provence, France

S. Nahar
Department of Building Materials and Structures, TNO, Molengraaffsingel 8, 2629 JD Delft, The Netherlands

N. Piérard
Belgian Road Research Centre, Woluwedal 42, 1200 Brussels, Belgium

G. Pipintakos
SuPAR, University of Antwerp, Groenenborgerlaan 179, 2020 Antwerp, Belgium

L. Porot
Kraton Polymers B.V., Transistorstraat 16, 1322 CE Almere, The Netherlands

K. Primerano
EMPA, Ueberlandstrasse 129, CH-8600 Dübendorf, Switzerland

A. Sharma · P. Tavassoti
Department of Civil and Environmental Engineering, University of Waterloo, 200 University Ave. West, Engineering 2 Building, Waterloo, ON N2L 3G1, Canada

S. Weigel
Department 7 Safety of Structures, Division 7.1 Building Materials, Federal Institute for Materials Research and Testing, Berlin, Germany

J. Wetekam
University of Kassel—Construction and Maintenance of Road Pavements, Mönchebergstraße 7, 34125 Kassel, Germany

J. Zhu
Swedish National Road and Transport Research Institute (VTI), Olaus Magnus väg 35, 581 95 Linköping, Sweden

vibrate, resulting in characteristic bands in the IR spectrum observed through FTIR spectroscopy. This makes it a useful technique since bitumen is a complex mixture of many different molecules. Hence, the method provides a simplified overview of the material, revealing relevant IR active functional groups.

Beitchman et al. [2] were among the first to apply FTIR spectroscopy on bituminous binders from roofing, which revealed one of the first FTIR spectra with all the important absorption bands, which are often reported in literature [3–7]. These bands can be categorized into hydrocarbon-containing groups like alkyls, alkenes or aromatic structures, oxygen-containing functional groups like ketones, 2-quinolones, carboxylic acids, anhydrides or alcohols [5, 7], and sulfur containing functional groups like sulfides, sulfoxides, sulfones or sulfate esters [8]. Ageing processes incorporate oxygen into the material, leading to increased absorption of two key functional groups: the carbonyl signal at 1700 cm^{-1} and the sulfoxide signal at 1030 cm^{-1}. Other less pronounced changes in the spectra can be seen in the spectral region between 1800 and 680 cm^{-1}, where an increase in absorption is linked to an increase in polarity due to the formation of higher polar fractions during ageing [5, 9]. Additionally, slight increases in the aromatic bands at 1600, 850, 810 and 750 cm^{-1} are expected upon ageing, which can be linked to the dehydrogenation of perhydro-aromatics like 9.10-dihydroantracenes. These molecules are believed to react with oxygen, initiating a chain oxidation mechanism [10].

In addition to its analytical capability, the practical applicability of the method needs to be highlighted. The development of the attenuated total reflection (ATR) mode, where the solid sample is brought in direct contact with the ATR crystal, simplifies the process and makes it easier to capture chemical information on bitumen without extensive knowledge or expertise. Furthermore, the method is time efficient since recording a spectrum only needs a couple of minutes. However, it should be noted that it may have some limitations due to variation in the resulting spectrum caused by differences in the detectors' characteristics or due to other bias caused by the ATR mode.

In the past, FTIR spectroscopy was not always an easily applicable and fast method. Before the implementation of the ATR mode, FTIR spectroscopy was performed in transmission mode [3, 7, 11]. Usually, this procedure required dissolving the binder in a solvent and applying it onto a suitable substrate, such as crystal window made from sodium chloride (NaCl), cesium iodide (CsI) or potassium bromide (KBr) and re-precipitating it [12]. Thus, the solvent needed time to evaporate before measurement, which involved significant time investment and risk of changes, as the thin bitumen film was exposed to the atmosphere. Other factors, such as dissolution rates and variations in solvents had to be considered, since the chemical composition of bitumen can differ depending on its crude oil source and refinery procedure. Alternatively, bitumen could also be measured in solution, which shows significantly higher intensities. However, the usage of solvent resembles a disadvantage in regard to practicability and work safety [13]. Transmission measurements could also be done without dissolution, by forming a thin film of hot binder on a crystal window. However, the question of repeatability and homogeneity can be raised. Therefore, a simple measurement with a solid material that is directly applied onto the ATR crystal can make practice easier. The most common type of crystal used in ATR mode is diamond, rarely these crystals are made out of germanium, zinc selenide or a combination of diamond and zinc selenide. The diamond crystal is often sintered into a metal plate, making the method robust and resistant to external stress factors such as mechanical force, solvents or different temperatures.

While the utilization of ATR-FTIR increased significantly over the last 10 years, it led to questions regarding the impact of sample preparation, measurements routine and universal applicability of the method: are we obtaining the same results when different laboratories are performing FTIR spectroscopy on the same material? The challenge lies in the fact that FTIR spectroscopy offers qualitative information about the presence of functional groups, which often needs to be referenced (e.g. comparing different ageing states of the same binder). This differs from analysis methods established in the engineering field, such as rheology or other mechanical tests, which provide specific quantified values linked to physical properties like stiffness. In addition, typical mechanical tests for bitumen are standardized. However, in the field of ATR-FTIR spectroscopy, various international standards exist, which are typically linked for polymer analysis like the determination of the microstructure

of styrene-butadiene rubber [14]. Limited availability of standardization of ATR-FTIR spectroscopy of bituminous binders cause different procedures adopted by different labs, resulting in inconsistencies in bitumen evaluation, i.e. different indices and values.

Typical spectral evaluation included normalization and integration of certain regions of interest, such as the carbonyl signal around 1700 cm^{-1} and the sulfoxides signal at 1030 cm^{-1}, which were linked to ageing and were quantified against unaged binders. Variations in the limits of integration have been proposed, such as the valley-to-valley method or tangential integration method, which considered the "actual" area of a band, excluding its background, proposed by Lamontagne et al. [3, 15]. The same methods, with slight adjustments to the limits, have been used by researchers from the Belgian Road Research Group (BRRC) [13], the Laboratoire central des ponts et chaussées (LCPC) [16], or participants from the MURE project in France [17]. Hofko et al. [4] compared this method with the full baseline integration method, addressing reproducibility and sensitivity to oxidation effects. Studies by Weigel et al. [12, 18] used standard normal variate (SNV) transformation and multivariate analysis (MVA) to detect differences in binder's crude oil origin and predict properties. Principal component analysis (PCA) and linear discriminant analysis (LDA) were used by Weigel et al. [12] and Ma et al. [19] to investigate the effects of ageing via FTIR spectroscopy. These methods can be used to analyze the entire spectrum easily and help identify differences in the bitumen source, type and ageing state. Similar work by Primerano et al. [20] used MVA to detect differences in laboratory and field-aged binders, addressing the influence of temperature, light and reactive oxygen species.

International groups, such as those within the RILEM community, have made use of FTIR spectroscopy. The RILEM TC 206-ATB on Advances in Interlaboratory Testing and Evaluation of Bituminous Materials has used FTIR spectroscopy and conducted spectral interpretations of international data [21], while the RILEM TC 252-CMB on Chemo-Mechanical Characterization of Bituminous Materials conducted a round robin test using FTIR spectroscopy. Nine different laboratories participated in the round robin and investigated four unaged, laboratory short- and long-term aged binders [22]. Data evaluation involved partial baseline correction and integration using either the tangential or full baseline method. These studies showed that a combination of baseline correction and full baseline integration led to better reproducibility compared to tangential integration. However, a major drawback of this round robin test was that each laboratory performed individual ageing procedures, which caused discrepancies in the ageing levels of the binders. The RILEM TC 272-PIM on Phase and Interface behavior of bituminous materials analyzed seven different binders, including unmodified and polymer-modified bitumen [23]. Results were evaluated using a deconvolution model based on a Gaussian normal distribution, derivate analysis, an 8-point baseline correction and maximum normalization at the aliphatic region around 1450 cm^{-1}, which showed promising comparisons between the laboratories. It also showed that the use of the derivate, by offsetting any baseline difference, offers a powerful tool to standardize the spectra. Mirwald et al. [24, 25] investigated the impact of heating time and temperature and storage conditions and time on unmodified and modified asphalt binders with ATR-FTIR spectroscopy. The results recommended rigorous specimen preparation, including a maximum heating time of 5 – 10 min below 180 °C (depending on the sample quantity and grade) and careful thermal monitoring and homogenization. Sample storage can have a significant effect since visible light can rapidly oxidize the sample surface, which alters the measured spectrum [25, 26]. As a result, binder samples should be stored in a dark, temperature-controlled room, covered with a non-light transparent lid, and measured within one hour after preparation to minimize environmental contamination from visible light, dust or elevated temperatures. These parameters are now considered in the current round robin test from the RILEM TC 295-FBB, with its objectives described below.

2 Objective and goals of the round robin test

This round robin test aimed to evaluate the impact of different sample preparation procedures for unmodified bituminous materials analyzed with ATR-FTIR spectroscopy, with an overview given in Fig. 1. A total of 21 active laboratories received three unaged, laboratory short-term and long-term aged binders and performed various preparation techniques.

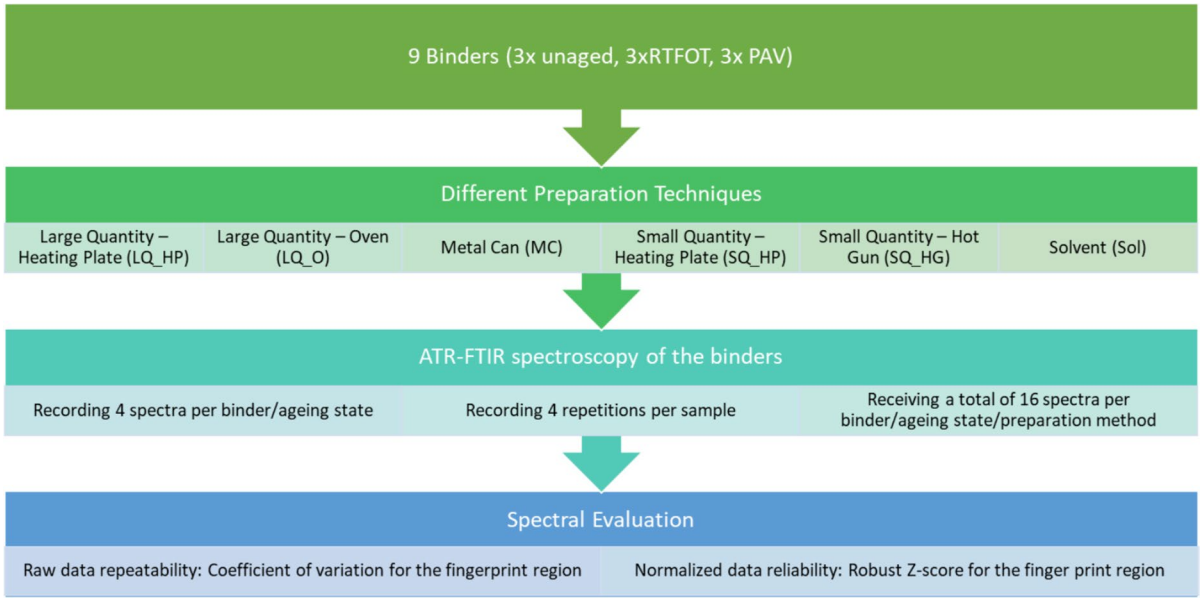

Fig. 1 Overview of the testing procedure in the RILEM TC 295-FBB TG1 round robin test and data evaluation methods

Four samples were prepared and measured with four repeats for each binder, ageing state and preparation method, resulting in a total of 16 spectra per binder, ageing state and preparation technique. The resulting raw data was collected and evaluated following a simple reproducibility evaluation (coefficient of variation) in the extended fingerprint region (spectral region between 1800 and 600 cm^{-1}) and a spectral evaluation approach to evaluate the consistency of the methods by identifying outliers with a high CV as well as outliers depending on the respective data collected per preparation method using a Python script.

2.1 Participating laboratories

Table 1 shows the 21 laboratories from 12 different countries working in academia or industry that participated in the round robin test.

3 Materials and methods

3.1 Materials

Three different unmodified 70/100 penetration graded binders from the same European specification class,

Table 1 Participating laboratories in the RILEM TC 295-FBB TG1 round robin test

Laboratory Name	Country
TU Wien	Austria
University of Antwerp + Nynas	Belgium
Belgian Road Research Centre	Belgium
University of Waterloo	Canada
Aalto University	Finland
Cerema / UMR MCD	France
Kraton	France
Colas	France
Bundeanstalt für Materialforschung und -prüfung (BAM)	Germany
Universität Kassel	Germany
Vilnius Techn	Lithuania
Road and Bridge Research Institute	Poland
Warsaw University of Technology	Poland
VTI	Sweden
EMPA	Switzerland
TNO	The Netherlands
TU Delft	The Netherlands
Ooms Producten	The Netherlands
The University of Texas at Austin	USA
FHWA	USA
MTE Services	USA

but different providers, were used in the round robin test. Their mechanical properties are as follows:

- B1158 (TU Wien): Needle penetration of 84 1/10 mm and softening point of 45.8 °C
- B1198 (TU Delft): Needle penetration of 85 1/10 mm and softening point of 47.0 °C
- B1199 (Nynas): Needle penetration of 84 1/10 mm and softening point of 45.8 °C

All binders were investigated at three different ageing states:

- Unaged/original (ageing suffix A)
- Laboratory short-term aged (ageing suffix B)
- Laboratory long-term aged (ageing suffix C)

Laboratory ageing was conducted at a single laboratory (TU Wien) to limit differences due to laboratory facilities, equipment and ageing protocols, ensuring that each laboratory received binders in the exact same ageing state. All aged binders were homogenized prior to distribution, flushed with nitrogen and sealed before shipping.

For laboratory short-term ageing, the Rolling Thin Film Oven Test (RTFOT) was performed according to the EN 12607-1 [27], with all binders aged at 163 °C for a duration of 75 min. For laboratory long-term ageing, the Pressure Ageing Vessel Test (PAV) was conducted according to the EN 14679 [28]. The ageing parameters were set to 100 °C and 2.1 MPa with an ageing duration of 20 h.

Overall, all active participants of the round robin test received a total of nine binders:

- B1158A (unaged), B1158B (RTFOT), B1158C (RTFOT + PAV)
- B1198A (unaged), B1198B (RTFOT), B1198C (RTFOT + PAV)
- B1199A (unaged), B1199B (RTFOT), B1199C (RTFOT + PAV)

3.2 Sample preparation methods

Six different sample preparation techniques, which involved utilizing various heating devices and tools commonly found in bitumen laboratories, such as heating plates, ovens, and hot guns, as well as different amounts of binder, were tested. Since FTIR spectroscopy can be conducted on a very small amount of sample, the preparation process considered a range of material quantities, from less than 1 g to more than 5 g of binder.

The purpose of using different sample preparation techniques was to determine whether significant differences between the methods can be detected based on exposure to thermal stress or other factors. An overview of the four preparation techniques that use common laboratory equipment is schematically shown in Fig. 2.

Since the major advantage of ATR-FTIR is that the material can be analysed in its solid state, most of the methods (five out of six) used the material in solid state. Only the last method involved dissolving the respective binders and re-precipitating them. This approach intended to determine whether the dissolving process affects the repeatability and reproducibility of the measurements. Furthermore, it aimed to provide a connection to previous measurements conducted in transmission mode [3, 7, 11].

An overview of the respective method parameters can be found in Table 2. Detailed insights into the individual preparation techniques will be elaborated below. If a method required a transfer of the binder from a larger can into a small metal can, a heated spoon or spatula was used. However, it was ensured that the spoon or spatula was not too hot (> 200 °C) to cause any evaporation of the binder.

3.2.1 Large quantity—heating plate (LQ_HP)

In the "Large Quantity—Heating Plate" (LQ_HP) method, ~ 5 g of the respective binder was transferred in a small, open metal container (see left side of Fig. 2) and was placed on a preheated heating plate that was set to 180 °C. During the heating process (5–10 min) the bitumen was continuously stirred with a thermometer to ensure sufficient homogeneity and maintain temperature control during the heating process. Once the binder had reached a sufficient workability or viscosity, the thermometer was used to prepare 4 bitumen droplets onto sufficient substrates (silicone paper or paper slips). The final samples were stored in a light and dust proof container and stored for a minimum of 5 min and a maximum of 1 h prior to the measurement.

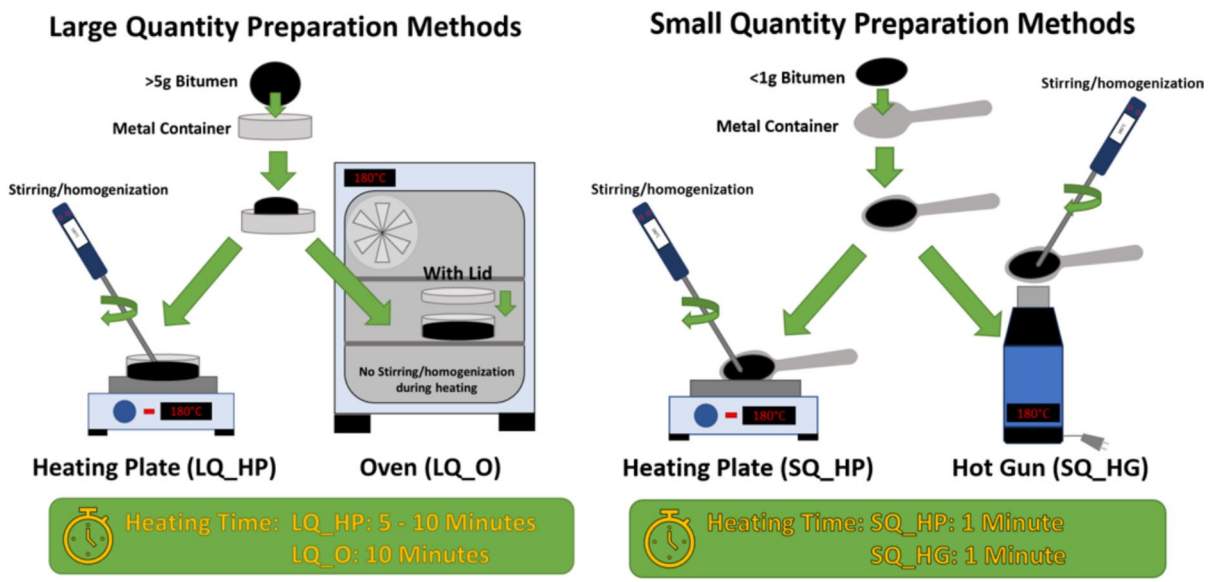

Fig. 2 Schematic drawing of the different solid sample preparation techniques involving common laboratory equipment

Table 2 Overview of the parameters for all sample preparation methods

Preparation method	Binder quantity [g]	Heating temperature [°C]	Heating duration [min]	Solvent [g]
Large Quantity—Heating Plate	> 5	180	5–10	–
Large Quantity—Oven	> 5	180	5–10	–
Metal Can	–	–	–	–
Small Quantity—Heating Plate	< 1	180	1–2	–
Small Quantity—Hot Gun	< 1	180	1–2	–
Solvent	1	–	–	3

3.2.2 Large quantity—oven (LQ_O)

In the "Large Quantity—Oven" (LQ_O) method, ~ 5 g of the respective binder was transferred in a small, closed metal container and was placed in a preheated oven set to 180 °C (see middle left side of Fig. 2). After 5–10 min of heating, the hot binder is taken out of the oven, stirred and homogenized with a thermometer and droplets are applied onto the substrate. Since the metal can is covered with a lid, ventilation settings of the oven were not significant [24]. The resulting samples were stored in a light and dust proof container and stored for a minimum of 5 min and a maximum of 1 h prior to the measurement.

3.2.3 Metal can (MC)

For the "metal can" preparation technique the binder was directly taken from the shipped binder can and applied onto the ATR crystal of the FTIR spectrometer without any preheating or pre-homogenization. This preparation method is the fastest but could cause uncertainties in repeatability and reproducibility, as no homogenization is conducted. Comparison between the "heating and homogenizing" and this preparation method will be discussed throughout the results.

3.2.4 Small quantity—heating plate (SQ_HP)

For the "Small Quantity- Heating Plate" method, merely 0.5–1 g of the respective binder was transferred into a small metal spoon and heated directly on a preheated heating plate set to 180 °C (see middle right side of Fig. 2). The binder was heated for roughly 1 min until it reached a liquid-like state. During heating continuous stirring with a thermometer ensured sufficient homogeneity and thermal monitoring. Once the binder has reached a liquid-like state the thermometer was used to apply the 4 small sample droplets onto sufficient substrates before the samples were covered in a light and dust proof container. The measurements were conducted between 5 and 60 min of resting.

3.2.5 Small quantity—hot gun (SQ_HG)

For the "Small Quantity—Hot Gun" method 0.5–1 g of the respective binder are transferred into a small metal spoon and heated directly over a hot gun (see right side of Fig. 2). The crucial aspect of this method is to ensure that the binder is not getting too hot (e.g. reducing the hot guns power) so that the binder gets in a liquid-like state without going beyond 180 °C. Therefore, the thermometer is used to track the temperature and ensure homogeneity. After heating (max. 1 min), the binder is then applied onto the substrates and stored in a light and dust proof container before being measured after 5 min of resting.

3.2.6 Solvent (Sol)

The last preparation method used in the round robin test was the solvent method, where 1 g of the respective binder was dissolved in 3 g of a toluene. After ensuring complete dissolution 2 -3 droplets of the solved binder are applied onto the ATR crystal (where a background of the clean crystal was recorded prior to the application). After the solvent has evaporated from the crystal the spectra are recorded. Completion of evaporation could be determined by the preview feature of the software, which allows in-situ monitoring of the current spectrum. The measurement was started once the bands of the solvent had disappeared from the preview spectrum.

3.3 Analysis method

3.3.1 ATR-FTIR spectrometer and parameters

Information on the different FTIR spectrometers that were used in this round robin test can be found in Table 3.

All spectra were recorded within a wavenumber range of 4000–600 cm^{-1}, a resolution of 4 cm^{-1} and 24 scans.[1] A background spectrum of the empty, clean ATR crystal was recorded prior to each application of a sample. After the background spectrum recording was completed, the sample was applied within one minute to reduce the risk of altering the background of the resulting spectrum due to potential changes in the surrounding atmosphere.

For each binder (B1158, B1198, B1199), ageing state (A, B, C) and preparation method (LQ_HP, LQ_O, MC, SQ_HP, SQ_HG, Sol) four samples were prepared and tested four times, resulting in a total of 16 spectra per binder, ageing state and preparation method. After recording the four repeats per sample, the ATR crystal was cleaned thoroughly using a non-toxic bitumen solvent (e.g. limonene), followed by a fast-evaporating alcohol (e.g. isopropanol) which ensured a clean ATR crystal prior to the next background recording.

All recorded spectra were gathered for visual pre-evaluation prior to data evaluation. Once all raw data was sorted, spectral data evaluation was started.

3.3.2 Raw data evaluation

The following routine describes the evaluation of a data set from one preparation method (e.g. LQ_HP) from one laboratory. If a laboratory has performed a preparation method, a total of 144 spectra (4 samples × 4 repeats × 3 binders × 3 ageing states) were generated and evaluated as follows:

- The 144 spectra in their raw data form were loaded into OPUS (FTIR software of Bruker) and

[1] It should be noted that due to differences in the way devices acquire scans, this number might change from device to device. The reference device for this determination was a Burker Alpha II.

Table 3 Information on the FTIR spectrometers used in the round robin test

Lab Code	IR Device	Device Mode	ATR Crystal	Detector	Pressure lever available
1	Perkin Elmer Spectrum 3	ATR	Diamond/ZnSe	DTGS	Yes
2	Thermo Scientific Nicolet iS50	ATR	Diamond	DLaTGS	Yes
3	Shimadzu	ATR	Diamond	DLATGS	Yes
4	Bruker Alpha II	ATR	Diamond	DTGS	Yes
5	Bruker Alpha	ATR	Germanium	DTGS	Yes
6	Bruker Alpha II	ATR,	Diamond	DTGS	Yes
7	Bruker Alpha	ATR,	Diamond	DTGS	Yes
8	Bruker Alpha II	ATR	Diamond	DTGS	Yes
9	Bruker Tensor 27 FTIR	ATR	Diamond	MTC	Yes
10	Thermo Scientific Nicolet iS10	ATR	Diamond	DTGS	Yes
11	Thermo Scientific IS10	ATR	Diamond	DTGS	No
12	Perkin Elmer	ATR	Diamond	LiTaO3	yes
13	Perkin Elmer spectrum 3	ATR	Diamond/ZnSe	LiTaO3	Yes
14	Perkin Elmer spectrum 3	ATR	Diamond/ZnSe	LiTaO3	No
15	Thermo Scientific Nicolett iS50	ATR	Diamond	DTGS	Yes
16	Thermo Scientific Nicolet iS10	ATR	Diamond	DTGS	Yes
17	Bruker Alpha II	ATR	Diamond	DTGS	Yes
18	Bruker Alpha II	ATR	Diamond	DTGS	Yes
19	Perkin Elmer spectrum 100	ATR	Diamond	DTGS	Yes
20	Thermo Scientific Nicolet iS50	ATR	Diamond	DTGS	Yes
21	Thermo Scientific Nicolet 6700	ATR	Diamond	DTGS	Yes

converted into 144 text files using a macro within the software.

- The 144 text files were condensed into an Excel file using a custom-made script that grouped the 16 spectra per binder and ageing into a total of nine tabs (one for each binder in each ageing state). Hence, one tab contains the following information:
  - 1st column contains the wavenumbers from 4000 to 600 cm^{-1}
  - 2nd—17th columns contain the absorbance values of the 16 spectra.
  - 18th, 19th and 20th columns contain the mean, standard deviation and coefficient of variation (standard deviation divided by mean) for each of the 16 repeats across the entire spectral range. Thus, a mean value, standard deviation and coefficient of variation are generated for each wavenumber from these 16 spectra.
- In addition to that, the mean, standard deviation and coefficient of variation from the so called extended fingerprint region, which refers to the spectral region between 1800 and 600 cm^{-1}, were collected, as they contain the most relevant spectral information for bituminous materials. Furthermore, this step neglects any scattering caused by effects like own absorption of the diamond crystal (between 2200 and 1800 cm^{-1}) or other environmental factors as well as ensures that absorption values are high enough to not lead to absurd values for the coefficient of variation.[2]
- The resulting coefficient of variation, abbreviated CV (in %) of the extended fingerprint region

[2] This can happen when absorption values are close to 0, where then a mean value close to 0 is divided by a standard deviation, resulting in coefficients of variation going beyond 100%, even though the reproducibility of the data itself is fine.

between 1800 and 600 cm^{-1} was summarized for each of the nine binders.

This procedure was carried out for all nine tabs and summarizes the data from each lab per ageing method, which provides information on the respective repeatability of one laboratory. This process was repeated for each laboratory evaluating the CV for the same preparation method. After completing this step, all CVs obtained from the labs were summarized into a master excel file, representing the reproducibility for the sample preparation method. This entire process was then repeated for all other preparation techniques. Finally, the resulting CV values are plotted as bar charts and are shown in the results from Figs. 3, 4, 5, 6, 7 and 8.

3.3.3 Robust Z-score analysis for outlier detection

In this study, a robust z-score analysis was employed to identify outliers within the dataset. This method offers advantages when handling datasets that may contain outliers capable of distorting traditional statistical measures. The analysis was conducted in two distinct manners: examining both positive and negative tails of the distribution and focusing solely on the positive tail. The positive and negative tails of a distribution refer to the regions where the values are significantly higher or lower than the mean, respectively, representing the extremes of the data range. In a normal distribution, these tails correspond to the right (positive tail) and left (negative tail) sides, where values are greater or lower than the mean by several standard deviations, respectively.

The study involving both tails of the distribution applied to spectral data aimed to detect spectra positioned significantly higher or lower compared to others in the dataset. This analysis provided insights into spectra that exhibited notable deviations from the norm. Conversely, the positive tail analysis focused on coefficient of variation (CV) values derived from multiple laboratories. This method targeted CV values showing unusually high variations, indicative of potential outliers within the dataset. Detailed explanations of both the both-tails and positive-tail approaches will be provided in the following sections. These analyses were influential in ensuring the robustness and consistency of the outlier detection process within diverse datasets.

3.3.3.1 Positive tail analysis This analysis focused solely on the positive tail, which used the CV values derived from multiple laboratories. This method targeted CV values showing unusually high variations, indicative of potential outliers within the dataset. The negative tail was not included since it represents a low CV value (good reproducibility). The analysis focused on CV values obtained from the spectral raw data evaluation. The goal was to identify outlier laboratories compared to others by applying the following procedure. The CV values from various laboratories, under identical conditions including preparation methods, binder type, and ageing level, were input into the robust z-score analysis algorithm using Eq. 1. This analysis facilitated the detection of laboratories exhibiting significant deviations from the expected variability, indicating potential outliers within the dataset.

$$\text{Robust Z} - \text{Score(xi)} = (\text{xi} - \text{Median(X)})/(\text{MAD(X)}) \quad (1)$$

In this study, Median(X) and MAD(X) refer to the median and median absolute deviation of the dataset X, respectively. MAD(X) is computed with Eq. 2:

$$\text{MAD(X)} = \text{Median}\left(|\text{X} - \text{Median(X)}|\right) \quad (2)$$

The dataset X corresponds to the individual values from each sample. In the context of the analysis, xi represents each laboratory's CV value.

A commonly used threshold is $\lambda = 3.5$, where data points with $|\text{Robust Z-Score(xi)}| > 3.5$ were classified as outliers [29, 30]. In a normally distributed dataset, about 99.954% of data points fall within ± 3.5 standard deviations from the mean. A z-score greater than 3.5 (or less than -3.5) is considered an outlier because it indicates a data point that is more than three standard deviations away from the mean. Consequently, only about 0.046% of data points are expected to have z-scores greater than 3.5 or less than -3.5. If significantly more than 0.046% of the points in each spectrum are outside the ± 3.5 z-score range, it could indicate that the entire dataset has characteristics of an outlier, suggesting a non-normal distribution or the presence of a broader issue with the spectrum (e.g., a major chemical change, baseline shift, or instrumental error).

Labs were flagged as outliers if their robust z-scores exceeded the predefined threshold λ. For this analysis focusing on the positive tail, outliers were

identified by Robust Z-Score(xi) > λ, where λ is typically set to 3.5 for standard outlier detection [29, 30].

3.3.3.2 Positive and negative tails analysis The data obtained from the round robin test was categorized based on various factors including the laboratory, FTIR device, preparation methods, binder source, or binder ageing condition. The primary focus of this study was to evaluate differences in FTIR measurements based on sample preparation methods. For this purpose, the data was initially divided into six groups representing different preparation methods: LQ_HP, LQ_O, MC, SQ_HP, SQ_HG, Sol.

The spectral data within each preparation method group underwent several preprocessing steps. First, a normalization procedure called normalization to change the maximum to one (NMO) was applied [31]. This normalization step aims to standardize absorbance values across data collected from different devices by scaling them to a consistent range of 0 to 1. Specifically, within the NMO method, the minimum absorbance value within the range of 2800–3200 cm^{-1} is set to zero, and then the maximum value is adjusted to 1 using Eqs. (3) and (4).

$$y_{ij}^{*} = y_{ij} - \min\left(y_i\right) \tag{3}$$

$$y_{ij}^{*} = \frac{y_{ij}}{\max\left(y_i\right)} \tag{4}$$

where yij and yij* represent the initial and new values of the j-th variable (i.e., absorbance) in the i-th spectrum (i.e., sample).

This preprocessing step is crucial in this phase because the data was measured using different devices, which can result in varying intensity levels across measurements. Normalization within each group ensures that the data becomes comparable without altering the inherent information of each spectrum. Following normalization, data trimming was performed to focus on the spectral data within the wavenumber range of 680–1800 cm^{-1}, aligning with the range used for coefficient of variation (CV) calculations.

By splitting the data into six groups based on preparation methods, the variability within each subgroup was increased, as each group contains three different materials and three different ageing levels. This approach allows to focus solely on the preparation methods. All three binder types and all ageing levels were aggregated to create an imaginary binder (superbinder) with a range of variability. This provides the opportunity to examine the consistency of results for each preparation method, independent of the binder type and ageing condition, and to identify spectra that deviate from acceptable norms. By interpreting the detected outliers, a general understanding of the outlier spectrum shape and perhaps the reasons for irregularities can be gained.

Within each subgroup, all absorbance values at a specific wavenumber were considered as X in Eq. 1, and robust z-scores were calculated for that wavenumber across all spectra. This process was repeated for all wavenumbers, resulting in each spectrum having a list of robust z-scores corresponding to its wavenumbers.

Outlier spectra were identified based on the criterion that if more than 1% of the points in the robust z-scores list exceeded a predetermined threshold λ.

4 Results and discussion

4.1 Raw data reproducibility and outlier detection

This section shows the raw data reproducibility as well as the impact of the positive tail outlier analysis and detection of all collected spectra. The bar figures show the respective CV from the extended fingerprint region between 1800 and 600 cm^{-1} of the 16 spectra for all 9 binders obtained from each lab (depending on whether the lab performed the method or not). The CV values of B1158 are shown in green, the CV values of B1198 in red and the CV values of B1199 in blue.

4.1.1 Large quantity—heating plate (LQ_HP) method

Figure 3 shows the resulting CV of the "Large Quantity—Heating Plate" method across the extended fingerprint region between 1800 and 600 cm^{-1}. The majority of the laboratories (17 out of 21) used this preparation method, as it was selected as the most commonly available technique in this round robin

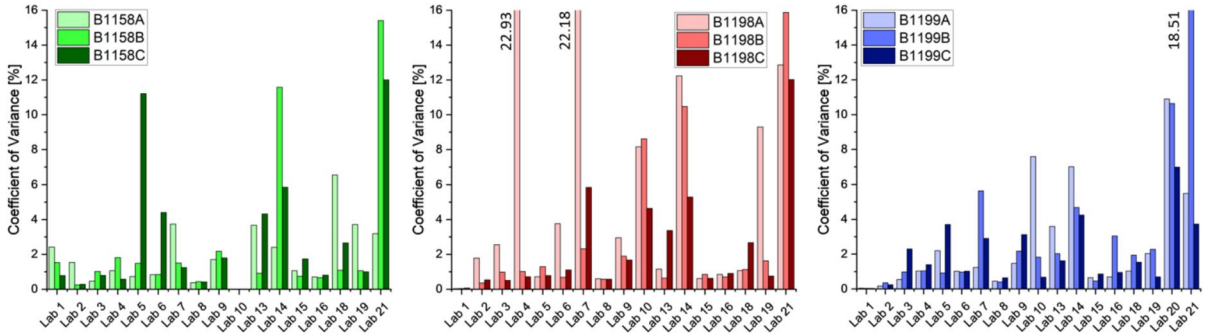

Fig. 3 Raw data reproducibility of the extended fingerprint region from all three binders in their respective unaged, RTFOT and PAV ageing states using the large quantity—heating plate (LQ_HP) method

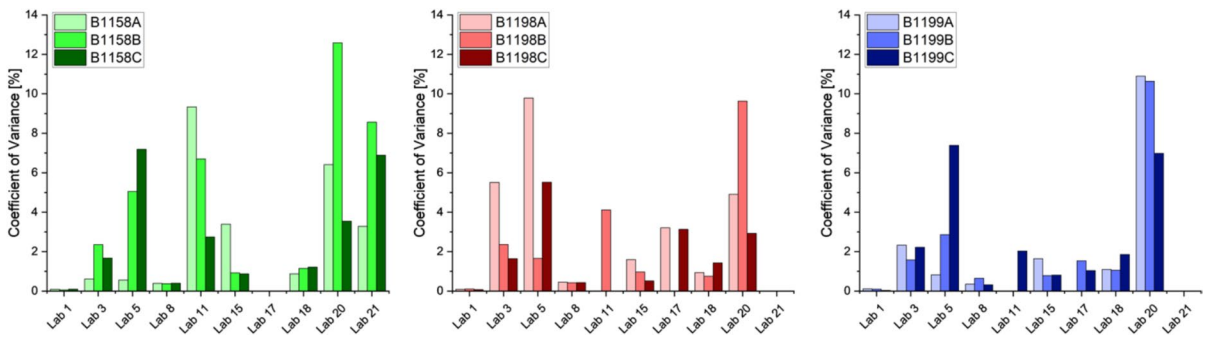

Fig. 4 Raw data reproducibility of the extended fingerprint region from all three binders in their respective unaged, RTFOT and PAV ageing states using the large quantity—oven (LQ_O) method

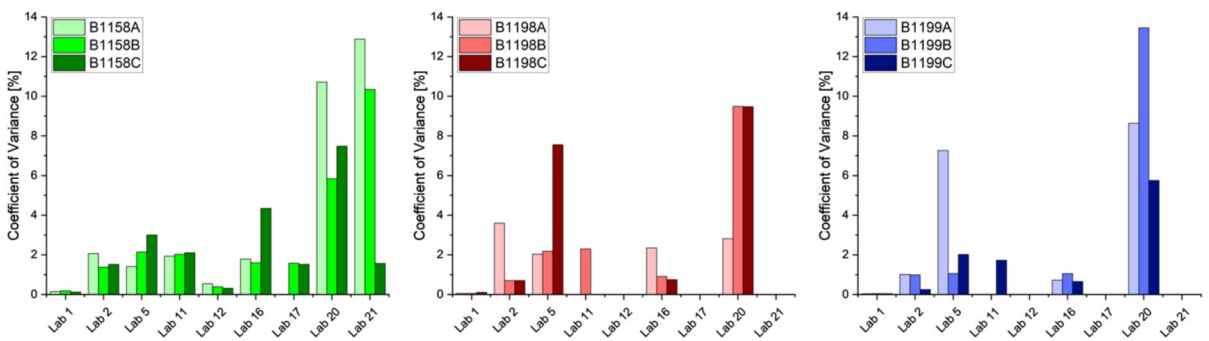

Fig. 5 Raw data reproducibility of the extended fingerprint region from all three binders in their respective unaged, RTFOT and PAV ageing states using the metal can (MC) method

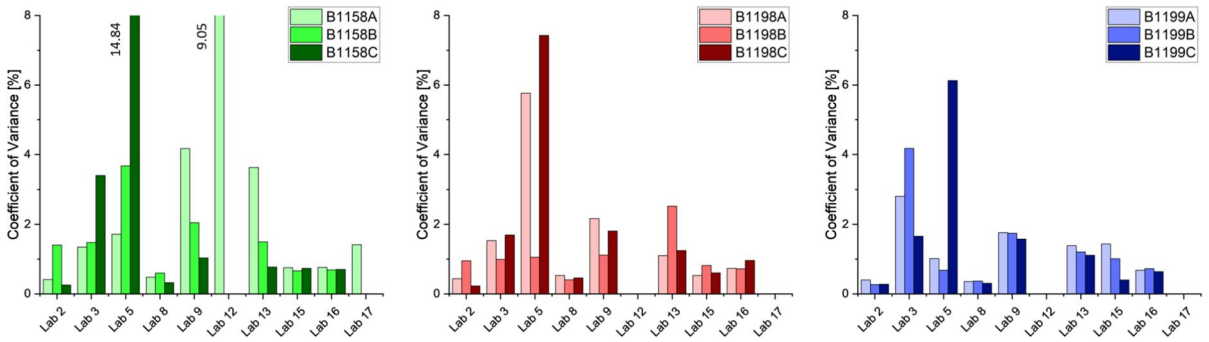

Fig. 6 Raw data reproducibility of the extended fingerprint region from all three binders in their respective unaged, RTFOT and PAV ageing states using the small quantity—heating plate (SQ_HP) method

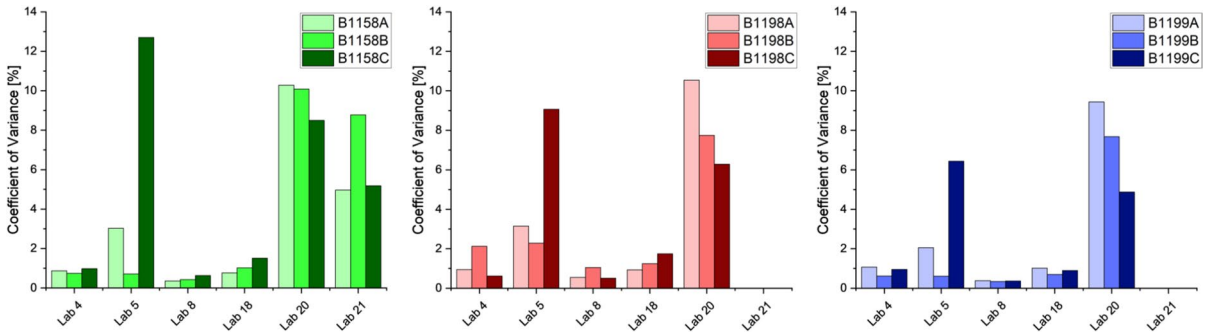

Fig. 7 Raw data reproducibility of the extended fingerprint region from all three binders in their respective unaged, RTFOT and PAV ageing states using the small quantity—hot gun (SQ_HG) method

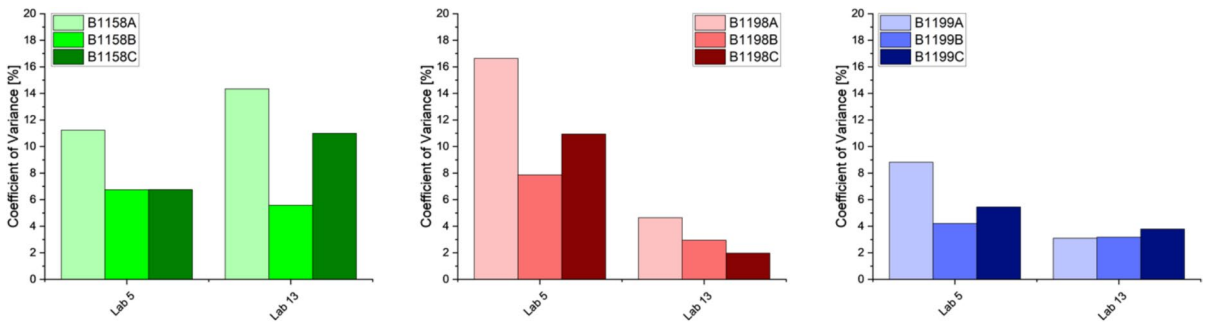

Fig. 8 Raw data reproducibility of the extended fingerprint region from all three binders in their respective unaged, RTFOT and PAV ageing states using the solvent (Sol) method

test. Across all three different binders, this method demonstrated overall good reproducibility. However, a wide range of CV values between 0.03% and 22.93% were observed. Summarizing all individual CV values into an "all binder" value, a CV of 2.89% can be obtained for the LQ_HP method.

Even when including the outliers, the results show the potential and great reproducibility of FTIR spectroscopy, where such a value was obtained from 2270 spectra recorded for this preparation method. From these 2270 spectra, 320 (20 samples) were detected as outliers during the positive tail analysis, meaning that the LQ_HP contains 14.1% outliers. Excluding these 320 spectra from the "all binder" results in a final CV value of 1.51%, which resembles an excellent reproducibility of the respective preparation method.

4.1.2 Large quantity—oven (LQ_O) method

Figure 4 shows the results of the raw data evaluation for the "Large Quantity—Oven" method. This method was performed by 10 laboratories resulting in a total of 1057 spectra considered in the evaluation process. The methods CV values range from 0.06 to 12.59%, which is a smaller span than for the LQ HP method. However, it is difficult to judge whether this method is more suitable for sample preparation than the LQ_HP by simply comparing the respective values obtained, as only 1057 spectra were considered, compared to the 2270 spectra for the LQ_HP method. This can mean that laboratories with a large outlier using the LQ_HP method might not have performed the LQ_O method.

Summarizing the CV values from each binder for all laboratories shown in Fig. 4, an "all binder" CV value of 2.76% is obtained. This value is slightly lower than the "all binder" value obtained for the LQ_HP method (2.89%). In order to judge which method is better for obtaining the best reproducibility, a comparison of the methods for one laboratory would be necessary. However, since the scope of this section is to discuss the raw data obtained in the round robin test, it would exceed the paper's content and will be discussed in future work, potentially in a recommendations document. A total of 272 spectra (17 samples) were detected in the positive tail outlier analysis. This means that 25.7% of the spectra recorded were outliers. Removing them lowers the "all binder" CV from 2.76 to 1.35%. This results in an even better reproducibility after outlier removal compared to the LQ_HP method.

4.1.3 Metal can (MC) method

Figure 5 shows the results of the metal can method. This method was carried out by 9 laboratories, resulting in a total of 780 spectra recorded. Values around the 2% mark can be seen for majority of the laboratories, with a CV range from 0.04 to 13.45%.

In contrast to the previously shown methods, this method does not utilize homogenization, as the sample is directly applied onto the ATR crystal. Thus, the question arises whether this could become a problem for reproducibility. However, no immediate difference can be seen in the "all binder" CV value of 2.95% compared to the values from the LQ_HP method (2.89%) or the LQ_O method (2.76%). However, it needs to be kept in mind that this method again only contains significantly fewer spectra (780) compared to the LQ_HP method (2270) and the LQ_O method (1057).

Out of the 780 spectra, a total of 96 spectra (6 samples) were detected as outliers during the positive tail analysis. This means that the MC method has an outlier percentage of 12.3%. The effect of removing these outliers reduces the "all binder" CV value to 2.01%. This ranks the method's reproducibility worse than the LQ_HP and LQ_O methods, which might indicate that homogenization could play a role in achieving the best reproducibility for FTIR spectroscopy.

4.1.4 Small quantity—heating plate (SQ_HP) method

Figure 6 shows the results of "Small Quantity—Heating Plate" method, where 15 laboratories have conducted measurements, resulting in a total of 1324 spectra. Most of the data has a CV of below 2%, with a range from 0.28 to 14.84%.

Summarizing all the CVs of all the labs and binders results to an "all binder" CV value of 1.68% for the SQ_HP method. Even with outliers, the method has a CV of below 2%. Removing the 80 outlier spectra (5 samples) during the positive tail analysis, the value was lowered even further to 1.18%. Considering this, the SQ_HP has an outlier percentage of 6.0%,

making it the method with the best reproducibility and lowest outlier rate compared to the metal can and large quantity methods.

4.1.5 Small quantity—hot gun (SQ_HG) method

Figure 7 shows the results of the "Small Quantity—Hot Gun" method, which was performed by six laboratories, resulting in a total of 770 spectra. CV values for this method range from 0.34 to 12.70%.

The respective "all binder" CV value is 3.24%, which becomes 1.91% after removing the 32 outlier spectra (2 samples—4.2% outliers) during the positive tail analysis. This ranks the method's reproducibility in 4th place after all other homogenizing preparation methods.

4.1.6 Solvent (Sol) method

Figure 8 shows the resulting CV values of the solvent method that was performed by two laboratories, resulting in a total of 260 spectra. Overall, the reproducibility seems to be worse compared to the non-solvent methods, as the lowest CV values can be found at 1.98%, reaching a maximum at 16.64%.

The previous assumption of worse reproducibility is confirmed by the "all binder" CV value of 7.18%. This is the highest value of all six preparation techniques, which indicates that the reproducibility is not reliable, in addition to the slow and tedious preparation procedure. Furthermore, since only two laboratories have performed the method and their values are all in the 5% range, no outliers are identified during the positive tail analysis, leaving the "all binder" CV after outlier identification at the same value. Nonetheless, it should be kept in mind that only two laboratories have performed this method. To fully judge its reliability and consistency, more data would be needed.

In addition to differences in reproducibility, there is a question regarding whether the dissolution process can influence the chemical functional groups detected with FTIR spectroscopy. However, as this paper focuses only on the raw data reproducibility of the methods used, addressing this concern will be part of future work.

4.2 Outlier identification and categorizing

To identify why certain spectra were assigned as outliers during the positive tail analysis, a manual check had to be conducted. Therefore, the respective spectra were loaded into the OPUS software and visually inspected. Most of the outliers could be separated into these two categories:

- One (or sometimes two) sample(s), consisting of four repeats or spectra showed designated differences when compared to the rest of the series (as demonstrated on the left side of Fig. 9).

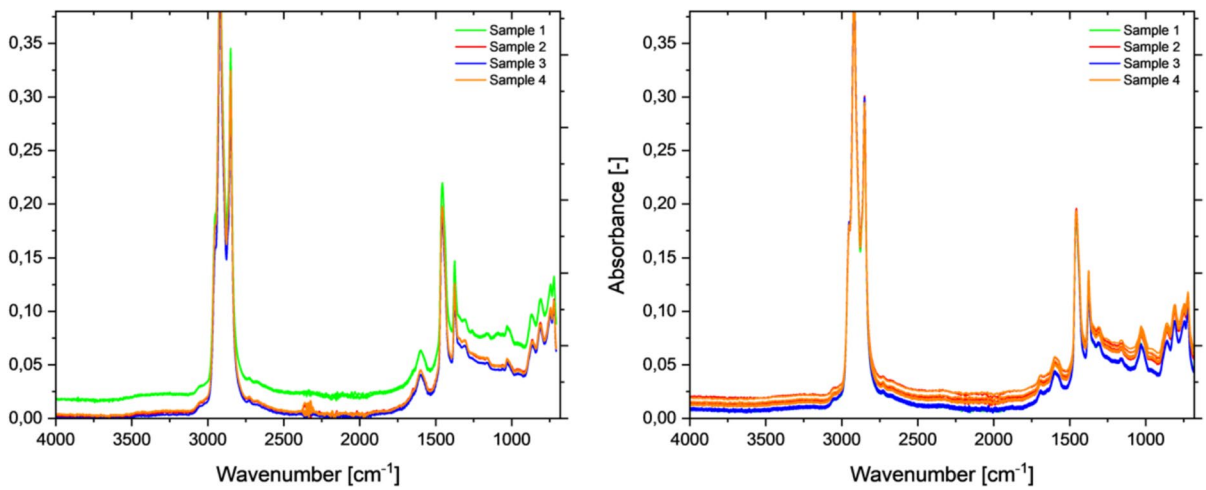

Fig. 9 Examples of common outliers from the round robin test

Table 4 Outlier analysis from the LQ_HP method

Method	Lab	Binder	Reason for outlier	Additional note
LQ_HP	4	B1198A	1 of 4 Samples is an Outlier	
LQ_HP	5	B1158C	1 of 4 Samples is an Outlier	
LQ_HP	7	B1198A	1 of 4 Samples is an Outlier	
LQ_HP	7	B1198C	1 of 4 Samples is an Outlier	
LQ_HP	10	B1198A	1 of 4 Samples is an Outlier	
LQ_HP	10	B1198B	1 of 4 Samples is an Outlier	
LQ_HP	10	B1199A	2 of 4 Samples are an Outlier	Significant Scattering between repeats
LQ_HP	14	B1158B	Different Intensities	
LQ_HP	14	B1158C	Different Intensities	
LQ_HP	14	B1198A	Different Intensities	
LQ_HP	14	B1199B	Different Intensities	
LQ_HP	14	B1199A	Different Intensities	
LQ_HP	18	B1158A	Different Intensities	Additional Bands (Impurities)
LQ_HP	19	B1198A	Different Intensities	
LQ_HP	21	B1158B	1 of 4 Samples is an Outlier	
LQ_HP	21	B1158C	Different Intensities	
LQ_HP	21	B1198A	1 of 4 Samples is an Outlier	
LQ_HP	21	B1198B	Different Intensities	
LQ_HP	21	B1198C	Different Intensities	
LQ_HP	21	B1199B	1 of 4 Samples is an Outlier	
LQ_O	5	B1158B	2 of 4 Samples are an Outlier	
LQ_O	5	B1158C	Different Intensities	
LQ_O	5	B1198A	2 of 4 Samples are an Outlier	
LQ_O	5	B1198C	1 of 4 Samples is an Outlier	
LQ_O	5	B1199C	Different Intensities	
LQ_O	11	B1158A	Different Intensities	
LQ_O	11	B1158B	Different Intensities	
LQ_O	20	B1158A	Different Intensities	
LQ_O	20	B1158B	Different Intensities	
LQ_O	20	B1198A	Different Intensities	
LQ_O	20	B1198B	Different Intensities	
LQ_O	20	B1199A	Different Intensities	
LQ_O	20	B1199B	Different Intensities	
LQ_O	20	B1199C	Different Intensities	
LQ_O	21	B1158B	1 of 4 Samples is an Outlier	
LQ_O	21	B1158C	Different Intensities	
MC	20	B1158A	Different Intensities	
MC	20	B1198B	Different Intensities	
MC	20	B1198C	Different Intensities	
MC	20	B1199B	Different Intensities	
MC	21	B1158A	Different Intensities	
MC	21	B1158B	Different Intensities	Additional Bands (Impurities)
SQ_HP	5	B1158C	Different Intensities	
SQ_HP	5	B1198A	Different Intensities	
SQ_HP	5	B1198C	2 of 4 Samples are an Outlier	
SQ_HP	12	B1158A	Different Intensities	
SQ_HG	5	B1158C	Different Intensities	

Table 4 (continued)

Method	Lab	Binder	Reason for outlier	Additional note
SQ_HG	5	B1198C	2 of 4 Samples are an Outlier	
SQ_HG	20	B1158A	Different Intensities	
SQ_HG	20	B1158B	Different Intensities	
SQ_HG	20	B1158C	Different Intensities	
SQ_HG	20	B1198A	Different Intensities	
SQ_HG	20	B1199A	Different Intensities	
SQ_HG	21	B1158B	Different Intensities	

- The entire series showed a difference in their overall absorbance (as demonstrated on the right side of Fig. 9), showing also scattering across the repeats within one sample.

Table 4 summarizes the outliers detected from the different preparation methods. From the table, it is apparent that almost all outliers are classified into specific categories. In rare cases, two of the four samples were around a similar intensity level. Additional notable occurrences in the spectra are detailed in the right column of the tables. Interestingly, only two samples showed a different band in their spectra, indicating that minimal error occurred during sample preparation as the occurrence of contaminated samples is low.

4.3 Consistency of the sample preparation methods

In addition to evaluating the reproducibility through coefficient of variation (CV) values, the consistency of inter-lab data for each preparation method, across all binders and ageing conditions was evaluated by aggregating the data into a "super-binder". To achieve this, the data were divided into six groups based on sample preparation methods. This aggregation allows to examine the variation in spectral measurements across all labs for the "super-binder" considering the impact of the different devices as

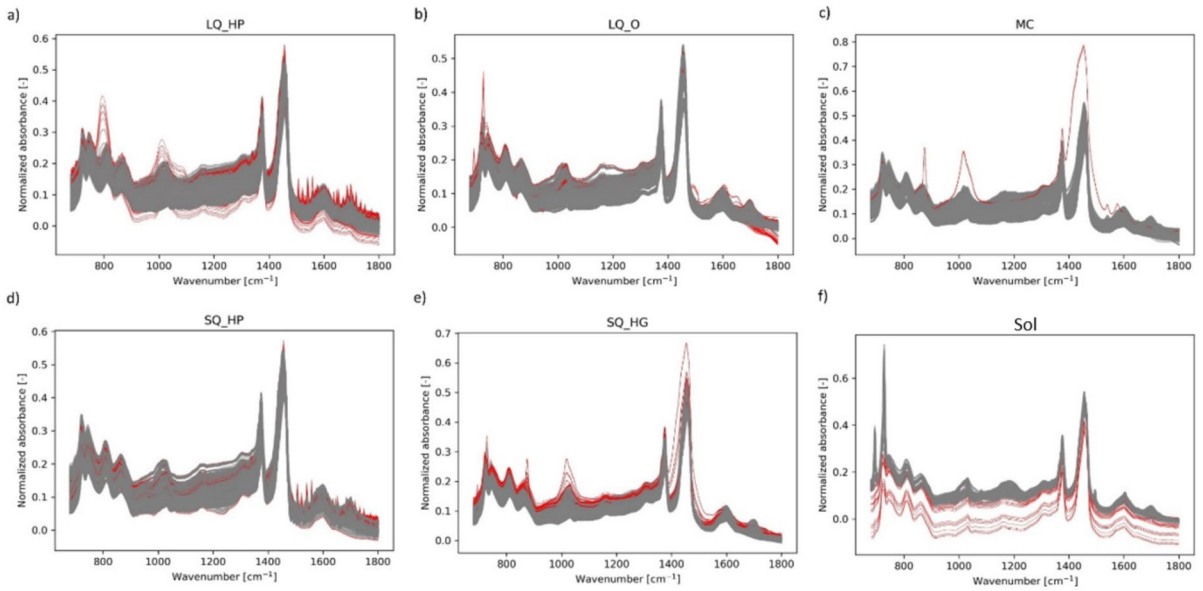

Fig. 10 Normalized ATR-FTIR spectra of all binders at all ageing levels plotted in the wavenumber range of 680–1800 cm^{-1} for each preparation methods: **a** LQ_HP, **b** LQ_O, **c** MC, **d** SQ_HP, **e** SQ_HG, **f** Sol. The detected outlier spectra are plotted in red, and the rest of the data is plotted in gray

well as the operator. Outlier detection was then applied to identify spectra that deviated from the normal range of measurements for the super binder, analysed both tails (positive and negative) as described in the methods.

It is important to note that being identified as outlier in the super-binder dataset does not necessarily indicate low reproducibility. Measurements with an acceptable CV value and good reproducibility may still be detected as outliers, if the cause of deviation (whatever it may be) is consistent across all measurements. On the contrary, there may be measurements where both low reproducibility and low consistency contribute to outlier detection.

Figure 10 a to 10 f show aggregated spectra from all laboratories for each preparation method, where red spectra are detected as outliers. The number of original and outlier spectra varies for each laboratory. To evaluate why certain spectra are assigned as outliers, a visual inspection was conducted. For better visualization, the respective outlier spectra are plotted in Fig. 11 a to 11 f. An outlier spectrum can belong to measurements from a specific laboratory without necessarily affecting all of its measurements. Since all measurements from all labs are combined, a wider range of variation is acceptable. In total, 8.4% out of 6461 were detected as outliers. Interestingly, the MC method demonstrated the lowest number of outliers, with only four spectra identified (out of 944 spectra). This would suggest that incorporating additional steps in sample preparation, without ensuring their proper execution, could pose challenges for future inter-laboratory comparisons. However, as the deviation in the spectra is most likely linked to the respective device used, a deeper look into this would be needed, as scattering in the lower end of the spectra is dependent on the spectrometer.

The following sections focus on identifying the characteristics of outlier spectra and understanding the reasons behind these deviations. This analysis aims to develop guidelines to enhance the effectiveness of sample preparation techniques.

The main reasons for considering a spectrum as an outlier are listed below for each preparation method:

- *LQ_HP*: Noise in the range of 1500–1800 cm^{-1}, differences in overall absorbance and baseline slope, unexpected peaks between 1000 and 1100 cm^{-1} and 1200–1400 cm^{-1}, and shifts in peak positions.

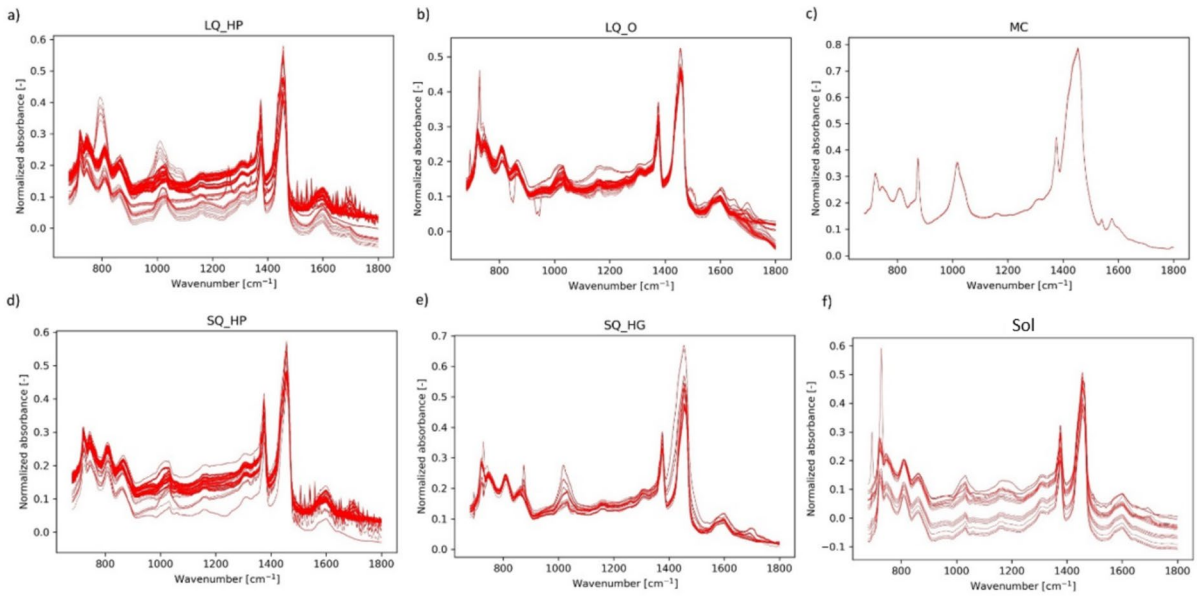

Fig. 11 Detected outlier normalized ATR-FTIR spectra of all binders at all ageing levels, plotted in the wavenumber range of 680–1800 cm^{-1} for each preparation methods: **a** LQ_HP, **b** LQ_O, **c** MC, **d** SQ_HP, **e** SQ_HG, **f** Sol

- *LQ_O*: Steeper baseline slope, abnormal valleys and peaks, and unexpected peaks between 1100 and 1300 cm^{-1} and 1400–1500 cm^{-1}.
- *MC*: Shifted peaks and unexpected absorbance for normal peaks at 833–912 cm^{-1} and 984–1047 cm^{-1}.
- *SQ_HP*: Noise in the range of 1500–1800 cm^{-1}, differences in overall absorbance and baseline slope, and unexpected peaks between 1300 and 1400 cm^{-1}.
- *SQ_HG*: Unexpected peaks between 680 and 900 cm^{-1} and shifted peaks towards higher absorbances.
- *Sol*: Noise in the range of 1500–1800 cm^{-1} and unexpected peaks between 680 and 750 cm^{-1}.

The observed issues can be grouped into three main categories: (a) differences in slope and baseline, (b) unexpected peaks and peak shifts, and (c) noisy spectra. These problems can result from the presence of impurities or contaminants in the bitumen sample, improper sample preparation (such as uneven spreading on the ATR crystal or variations in sample thickness), and misalignment or calibration errors of devices. To address these issues and obtain reliable spectra in ATR-FTIR spectroscopy of binders, consistent and proper sample preparation is crucial. Training of ATR-FTIR users to follow a commonly accepted and applied preparation method can eliminate most of these problems. In addition to proper sample preparation, baseline correction preprocessing, regular calibration of the instrument to ensure accurate measurements, increasing the number of scans, and performing signal averaging to improve the signal-to-noise ratio and reduce noise in the spectrum are recommended. Furthermore, if unexpected peaks or shifts are observed, repeating the measurements to confirm the results is suggested.

5 Summary and conclusion

The manuscript presents the results from the ATR-FTIR spectroscopy round robin test conducted in the Task Group 1 of the RILEM TC-295 Fingerprinting Bituminous Binders, where 21 laboratories measured three different binders in three different ageing states (unaged, laboratory short-term aged and laboratory long-term aged). 16 spectra were recorded per binder, ageing state and sample preparation method, which resulted in a total of 6461 spectra collected in the round robin test. The reproducibility was evaluated using the mean, standard deviation and coefficient of variation of the absorbance values from the extended fingerprint region (spectral range between 1800 and 600 cm^{-1}), followed by various outlier detection procedures, which lead to the following conclusions:

- The best overall reproducibility after outlier removal is seen for the "Small Quantity—Heating Plate" method (CV of 1.1% considering 1324 spectra), followed by the "Large Quantity—Oven" method (CV of 1.35% considering 1057 spectra), the "Large Quantity—Heating Plate" (CV of 1.51% considering 2270 spectra), "Small Quantity—Hot Gun" (CV of 1.91% considering 770 spectra) and the "Metal Can" method (CV of 2.01% considering 780 spectra). The method with the worst reproducibility is the solvent method (CV of 7.18% considering 260 spectra).
- The outlier detection was performed in two different forms: focusing on both the positive and negative tails, and exclusively on the positive tail. The first approach, applied to spectral data, identified spectra with values significantly higher or lower than the norm, providing insights into notable deviations caused mostly by the devices (detector type and its absorption profile) as well as the operator. This outlier identification and categorizing using the positive tail analysis revealed that most of the outliers could be clustered into two groups: in the first outlier group, one (or two) of the four samples showed a significant difference in absorption. In the second outlier group, a difference in absorption across all 16 spectra was observed.
- The second approach, focused on CV values from multiple laboratories, targeted outliers with unusually high variability, indicating potential inconsistencies from the respective laboratory. The results suggest that incorporating additional steps in sample preparation, without ensuring their proper execution, could pose challenges for future interlaboratory comparisons. Thus, a focus on proper execution is crucial.

Overall, the results from the evaluation of the reproducibility and consistency from the round robin test demonstrate the excellent capabilities of

ATR-FTIR spectroscopy. Following specific sample preparation techniques can yield excellent reproducibility, bringing the method one step closer towards pre-standardization. As the topic of the reproducibility of the method itself has been covered, future work will focus on questions related to whether a universal data evaluation technique can be found independent of a spectrometer. This could provide insights on specific questions, such as universal evaluation of binder ageing via indices generation. Only after this, a clear recommendation as to the most suitable sample preparation technique can be given.

Acknowledgements The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged. Furthermore, the authors would also like to express their gratitude to the CD laboratories company partners BMI Group, OMV Downstream and Pittel+Brausewetter for their financial support. This paper/article is created under the research program Knowledge-based Pavement Engineering (KPE). KPE is a cooperation between Rijkswaterstaat, TNO and TU Delft in which scientific and applied knowledge is gained about asphalt pavements and which contributes to the aim of Rijkswaterstaat to be completely climate neutral and to work according to the circular principle by 2030. The opinions expressed in this paper is solely from the authors. The partial support of the Natural Sciences and Engineering Research Council of Canada (NSERC) and Discovery Grant program RGPIN-2023-03727 is acknowledged. The authors would also like to thank Hilde Soenen from Nynas for providing one of the three binders used in the round robin testing as well as the actively participating laboratories from Colas (Xavier Cabonneau), MTE Services (Gerald Reinke), Ooms Producten bv (Kees Plug) and Peter Mikhailenko (formerly EMPA), who participated in the round robin testing.

Funding Open access funding provided by TU Wien (TUW). Christian Doppler Forschungsgesellschaft, 1836999, Bernhard Hofko, Rijkswaterstaat, 31164321, Aikaterini Varveri.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Petersen JC (1984) Chemical composition of asphalt as related to asphalt durability: state of the art. Transp Res Rec 999:13–30
2. Beitchman BD (1959) Infrared spectra of asphalts. J Res Nat Bureau Stand Sect a-Phys Chem 63(2):189–193
3. Lamontagne J (2001) Comparison by Fourier transform infrared (FTIR) spectroscopy of different ageing techniques: application to road bitumens. Fuel 80(4):483–488
4. Hofko B et al (2017) Repeatability and sensitivity of FTIR ATR spectral analysis methods for bituminous binders. Mater Struct 50(3):1–15
5. Mirwald J et al (2020) Investigating bitumen long-term-ageing in the laboratory by spectroscopic analysis of the SARA fractions. Constr Build Mater 258:119577
6. Petersen J, Barbour F, Dorrence S (1975) Identification of dicarboxylic anhydrides in oxidized asphalts. Anal Chem 47(1):107–111
7. Petersen JC et al (2002) Molecular interactions of asphalt. Tentative identification of 2-quinolones in asphalt and their interaction with carboxylic acids present. Anal Chem 43(11):1491–1496
8. Mirwald J, Werkovits S, Camargo I, Maschauer D, Hofko B, & Grothe H (2020) Time and storage dependent effects of bitumen—comparison of surface and bulk. In: RILEM International Symposium on Bituminous Materials (pp. 1853-1859). Cham: Springer International Publishing
9. Mirwald J et al (2020) Understanding bitumen ageing by investigation of its polarity fractions. Constr Build Mater 250:118809
10. Petersen J C (2009) A review of the fundamentals of asphalt oxidation: chemical, physicochemical, physical property, and durability relationships. Transportation research circular, (E-C140).
11. Petersen JC (1975) Quantitative method using differential infrared spectrometry for the determination of compound types absorbing in the carbonyl region in asphalts. Anal Chem 47(1):112–117
12. Weigel S, Stephan D (2018) Differentiation of bitumen according to the refinery and ageing state based on FTIR spectroscopy and multivariate analysis methods. Mater Struct 51(5):1–11
13. Pierard N (2013) Bitumen analysis with FTIR spectrometry: processing of FTIR spectra. BRRC protocol, ME85/13.
14. ISO, ISO 21561–2:2024 - Styrene-butadiene rubber (SBR) — Determination of the microstructure of solution-polymerized SBR — Part 2: Fourier transform infrared spectrometry (FTIR) with attenuated total reflection (ATR) method. 2024.
15. Lamontagne J et al (2001) Direct and continuous methodological approach to study the ageing of fossil organic material by infrared microspectrometry imaging:

application to polymer modified bitumen. Anal Chim Acta 444(2):241–250
16. Mouillet V, Farcas F, Battaglia V, Besson S, Petiteau C, & Lecunff F (2009) Identification and quantification of bituminous binder's oxygenated species. Analysis by Fourier transform infrared spectroscopy, Méthode d'essai LPC, 69
17. Dony A, Ziyani L, Drouadaine I, Pouget S, Faucon-Dumont S, Simard D, & Gueit C (2016) MURE National Project: FTIR spectroscopy study to assess ageing of asphalt mixtures. In: Proceedings of the E&E congress.
18. Weigel S, Stephan D (2017) The prediction of bitumen properties based on FTIR and multivariate analysis methods. Fuel 208:655–661
19. Ma L et al (2023) Chemical characterisation of bitumen type and ageing state based on FTIR spectroscopy and discriminant analysis integrated with variable selection methods. Road Mater Pavement Des 24(sup1):506–520
20. Primerano K et al (2023) Characterization of long-term aged bitumen with FTIR spectroscopy and multivariate analysis methods. Constr Build Mater 409:133956
21. Partl M et al. (2013) RILEM State-of-the-Art Reports
22. Hofko B et al (2018) FTIR spectral analysis of bituminous binders: reproducibility and impact of ageing temperature. Mater Struct 51(2):1–16
23. Porot L et al. (2020) Complex bituminous binders, are current test methods suitable for? in ISBM. 2020. Lyon.
24. Mirwald J, Nura D, Hofko B (2022) Recommendations for handling bitumen prior to FTIR spectroscopy. Mater Struct 55(2):26
25. Mirwald J et al. (2022) Time and storage dependent effects of bitumen—comparison of surface and bulk. In: Proceedings of the RILEM International Symposium on Bituminous Materials. 2022. Cham: Springer International Publishing.
26. Mirwald J et al (2022) Impact of UV–Vis light on the oxidation of bitumen in correlation to solar spectral irradiance data. Constr Build Mater 316:125816
27. CEN, EN 12607–1: Bitumen and bituminous binders - Determination of the resistance to hardening under the influence of heat and air - Part 1: RTFOT method. 2015: Brussels.
28. CEN, EN 14769: Bitumen and bituminous binders - Accelerated long-term ageing conditioning by a Pressure Ageing Vessel (PAV). 2012: Brussels.
29. Iglewicz B and D.C. Hoaglin, Volume 16: how to detect and handle outliers. 1993: Quality Press.
30. Dastjerdy B, Saeidi A, Heidarzadeh S (2023) Review of applicable outlier detection methods to treat geomechanical data. Geotechnics 3(2):375–396
31. Khalighi S et al (2024) Evaluating the impact of data preprocessing methods on classification of ATR-FTIR spectra of bituminous binders. Fuel 376:132701