Forensic Sciences Research, 2023, **8**, 313–320
https://doi.org/10.1093/fsr/owad045
Advance access publication date 19 December 2023
Research Article

Comparing baseline correction algorithms in discriminating brownish soils from five proximity locations based on UPLC and PLS-DA methods

Muhamad Adib bin Ahmad[1], Loong Chuen Lee[1,2,*], Nur Ain Najihah Mohd Rosdi[1], Nadirah Binti Abd Hamid[1], Ab Aziz Ishak[1], Hukil Sino[1]

[1]Forensic Science Program, CODTIS, Faculty of Health Sciences, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia
[2]Institute of IR 4.0, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia

*Corresponding author. E-mail: lc_lee@ukm.edu.my

Abstract

Soil is commonly collected from an outdoor crime scene, and thus it is helpful in linking a suspect and a victim to a crime scene. The chemical profiles of soils can be acquired *via* chemical instruments such as Ultra-Performance Liquid Chromatography (UPLC). However, the UPLC chromatogram often interferes with an unstable baseline. In this paper, we compared the performance of five baseline correction (BC) algorithms, i.e. asymmetric least squares (AsLS), fill peak, iterative restricted least squares, median window (MW), and modified polynomial fitting, in discriminating 30 chromatograms of brownish soils by five locations of origin, i.e. PP, HK, KU, BL, and KB. The performances of the preprocessed sub-datasets were first visually inspected through the mean chromatograms and then further explored *via* scores plots of principal component analysis (PCA). Eventually, the predictive performances of the partial least squares-discriminant analysis (PLS-DA) models estimated from 1 000 pairs of training and testing samples (i.e. prepared *via* iterative random resampling split at 75:25) were studied to identify the best BC method. Mean raw chromatograms of the 10 soil samples were different from each other, with evident fluctuated baselines. AsLS and MW corrected chromatograms demonstrated the most significant improvement compared with the raw counterpart. Meanwhile, the scores plot of PCA revealed that most of the sub-datasets produced three separate clusters. Then, the sub-datasets were modelled *via* the PLS-DA technique. MW emerged as the excellent BC method based on the mean prediction accuracy estimated using 1 000 pairs of training and testing samples. In conclusion, MW outperformed the other BC methods in correcting the UPLC data of soil.

Key points

- UPLC data of soil interfere with baseline drifts.
- BC can improve the quality of the pixel-level UPLC data.
- MW emerges as the most desired algorithm in improving the quality of UPLC data of soil.

Keywords: soil forensics; ultra-performance liquid chromatography; baseline correction; principal component analysis; partial least squares-discriminant analysis

Introduction

Soil comprises various components such as organic compounds, minerals, and inorganic compounds [1]. The soil composition is naturally affected by living organisms or altered by human activities, e.g. agriculture and mining [2]. Therefore, soils from different locations potentially show unique compositions and thus could be highly individualistic. In the context of forensic analysis, the soil is one of the most common types of trace evidence that can link a suspect/victim to a crime scene [3]. This is because soil can be easily transferred from one place to another and could be a reliable source to identify a particular location.

Most forensic studies have been concerned with seeking the best way to discriminate soils originating from different locations [4]. For example, Xu et al. [5] recently proposed using laser-induced breakdown spectroscopy and Fourier transform infrared total attenuated reflectance spectroscopy to characterize the soils according to the elements and organic compounds. On the other hand, Profumo et al. [6] demonstrated the benefits of gas chromatography-mass spectrometer (GC–MS) in forensic soil analysis by focusing on the volatile, semivolatile and volatilizable fractions of soil samples. Meanwhile, McCulloch et al. [7] found that the high-performance liquid chromatography (HPLC) technique could be feasible to differentiate soils from proximity locations.

It is worth noting that forensic analysts seldom place sufficient attention to preprocessing the data before interpretation. For instance, McCulloch et al. [7] employed high-performance liquid chromatography with diode-array detection (HPLC-DAD) and GC methods to discriminate geoforensic trace material from close proximity locations. Instead of pixel-level data, the authors integrated selected peaks from the HPLC-DAD and GC systems. Hence, the good performances of the statistical prediction models were attributed to the

Received: June 14, 2022. **Accepted:** September 26, 2023

pre-selection of peaks performed manually by the authors. However, it is generally agreed that pixel-level data are far more suitable than targeted peak table data in constructing an automatic data analysis pipeline [8].

Nonetheless, it is well known that pixel-level data comprise both the samples and irrelevant signals attributed to the inherent limitations of the instruments, e.g. instrumental drift and aged columns [9–12]. Therefore, this work aims to evaluate five baseline correction (BC) algorithms in improving the baselines of 30 Ultra-Performance Liquid Chromatography (UPLC) pixel-level data of soils [13–15]. The purpose was to discriminate the soils according to five locations of origin in Kajang district, Selangor, Malaysia, based on the UPLC fingerprints.

Materials and methods

Soil sampling

Five proximity locations in the Kajang district, Selangor, Malaysia, as detailed in Figure 1, were chosen to collect 10 soil samples. All the locations are accessible by the public without any permit (Supplementary Figure S1). Three of the five locations are abandoned land without monitoring by any authority, i.e. BL, KU, and KB. Meanwhile, the forest (HK) and Fernarium (PP) located on the campus of Universiti Kebangsaan Malaysia (UKM) are monitored by the campus security team. Two brownish soils, i.e. brown and yellowish brown, were collected from each location using the grid-pattern procedure described by Pye [1]. A stainless-steel spatula was used to collect ∼5 g of topsoil (0–1 cm depth) accumulated from the four corners and the central point of a 1 m square grid.

Preparation of soils extracts

All samples were kept in zip-lock plastic bags and placed in paper boxes to be air-dried overnight at room temperature on the same day of sampling. On the next day, the samples were further dried in an oven at 60°C for 3 h. After leaving cool in a desiccator, the samples were ground using a mortar and pestle, which was then sieved using a stainless-steel analytical sieve (600 μm). Then, ∼0.5 g of soil was placed into a 1.5-mL microcentrifuge tube and dissolved with ∼1 mL of HPLC-grade acetonitrile (ACN) (Fisher Chemical, Mumbai, Maharashtra, India). Following that, the snap cap of the tube was closed tightly before placing it into a sonicator for 20 min. Eventually, the tubes were centrifuged at 13 000 rpm for 15 min. The supernatant was transferred into a syringe locked to 0.2-μm polytetrafluoroethylene for filtering into an HPLC vial [16].

UPLC analysis

The UPLC analysis was performed by using a Waters ACQUITY UPLC™ system (Waters, Milford, MA, USA) equipped with a binary solvent manager, autosampler and photodiode-array detector (PDA). A Waters ACQUITY UPLCTM BEH C18 column (2.1 mm × 50 mm, 1.7 mm particle size) from Waters was chosen. The samples were separated in the column using isocratic elution with 90% ACN in water (containing 10% ACN); flow rate: 0.2 mL/min; run time: 15 min. The column temperature was set at 25°C, and the injection volume was 7 uL. The detection wavelength was set at 230 nm [16]. By performing triplicate injections

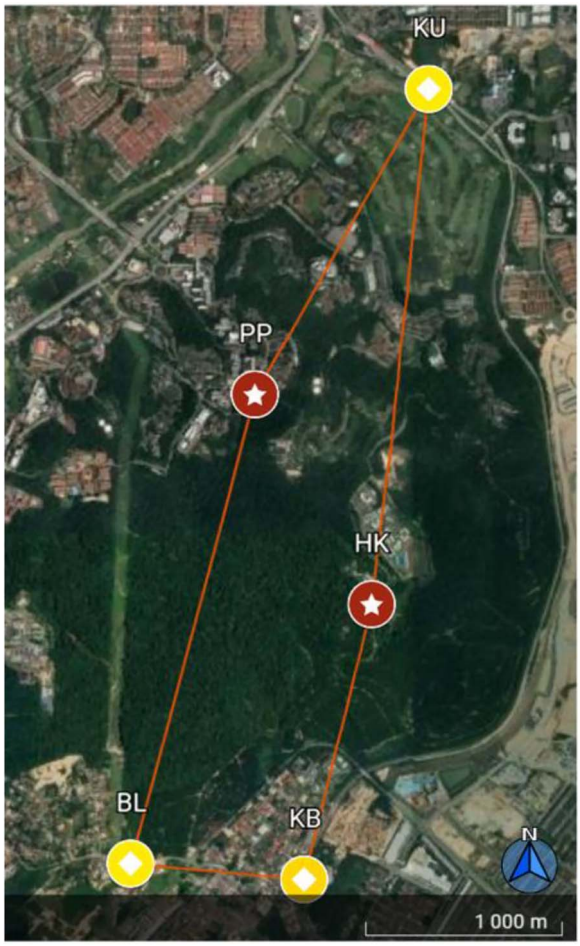

Figure 1 Location map of soil sample collection sites. Image from Google Earth 2022 (viewed 6 October 2022). KB: abandoned land nearby the Bangi commuter station (2.9008074″N, 101.7850107″E); BL: illegal trash dumping site (2.9015417″N, 101.7769922″E); KU: abandoned land nearby the Universiti Kebangsaan Malaysia (UKM) commuter station (2.9373368″N, 101.7907547″E); HK: UKM forest (2.9135556″N, 101.788083″E); PP: Fernarium UKM (2.9232222″N, 101.782722″E).

per vial, a total of 30 chromatograms were prepared from the 10 soil samples. Prior to statistical analysis, the pixel-level chromatographic data were constructed into a data matrix with 30 rows (i.e. UPLC chromatograms) and 18 000 columns (i.e. retention time, RT points). Based on preliminary inspection, it was found that the window after 5 min showed a minimal number of trivial peaks. Hence, it was decided that only the sub-RT window covering 1–5 min (i.e. 4 800 RT points) was studied.

Statistical evaluation

Five variants of BC algorithms available in the R package "baseline" [17], as listed in Table 1, employ a different principle in correcting the baseline of the chromatograms. An additional five treated sub-dataset were created by preprocessing the sub-RT dataset with the five BC algorithms. The impact of the BC algorithms was evaluated using three approaches: (i) visual inspection of the mean chromatograms; (ii) spatial distribution of samples in scores plot of principal component analysis (PCA); and (iii) predictive performance of partial least squares-discriminant analysis (PLS-DA) models. All the

statistical analyses were accomplished in the *R* statistical software, v. 3.6.2 [18].

As mentioned above, each chromatogram presented 18 000 variables; thus, the chromatographic data are of high dimensionality. Therefore, after comparing the treated mean chromatograms with the raw counterpart, the data were reduced using PCA. The most discriminative scores plot was identified by inspecting ten 2-dimensional scores plots deriving from the first five PCs. PLS-DA modelling [19] was also applied to both the raw data and treated counterparts to gain more insights into the discriminative capability of the data. PLS-DA deploying the naïve decision rule was employed to predict the location of origin of soil based on the pixel-level UPLC data. The data were first randomly split into 75% training and 25% testing samples repeated 1 000 times. Then, the model built using the training samples was tested using the corresponding testing samples. The mean predictive accuracy rate was estimated from the 1 000 pairs of training and testing samples. Given a sub-dataset, four models were incrementally constructed by considering the first five PLS components, i.e. PLS1-2, PLS1-3, PLS1-4, and PLS1-5.

Results

Mean chromatograms

Figure 2 presents the mean chromatograms of the 10 soil samples averaged by the five locations of origin deriving from the raw data and the five baseline corrected counterparts.

Referring to the raw chromatograms alone (Figure 2A), most peaks were unresolved and overlapped with proximity peaks, partly caused by the unstable baseline. The most undesired baseline was seen in the BL sample; the part of the baseline crowded by majority peaks immensely fluctuated. Meanwhile, the KU sample showed the least number of prominent peaks, denoting composed of the least number of non-volatile organic compounds than the remaining samples. The PP and HK were highly similar in terms of the UPLC fingerprint. This could be explained by the fact that both are located on the campus of UKM. Lastly, KB is readily identified according to the prominent peak eluted ~1.7–1.8 min.

At first glance, the most desired baseline was seen in asymmetric least squares (AsLS) treated chromatograms (Figure 2B). The AsLS algorithm has modified the overall chromatographic patterns, particularly the chromatogram of BL. Meanwhile, iterative restricted least squares (iRLS) (Figure 2D) and median window (MW) (Figure 2E) seemed to preserve the chromatographic pattern of BL whilst minimizing the baseline drift though not as good as achieved *via* AsLS. In particular, MW slightly outperformed the iRLS as the latter showed a more fluctuated baseline in 1–1.5 min. Last but not least, the other two algorithms, i.e. fill peak (FP) and modified polynomial fitting (MPF), have not caused any improvement to the raw data. (Figure 2C and F)

Thus, the relative performances of the MW and AsLS were thus further elucidated based on the scores plot of PCA and PLS-DA predictive modelling. In addition, the individual replicate chromatograms of the raw, AsLS and MW treated sub-datasets are presented in Supplementary Figures S2 and S3.

Scores plots of PCA

Next, the classification ability of the sub-dataset by the five locations of origin was assessed through the spatial clustering seen in the scores plot of PCA. After inspecting the 10 scores plots constructed with the first five PCs, it was found that only the PC1 *vs*. PC2 and PC2 *vs*. PC3 produced meaningful separation. The remaining plots mainly showed all the samples scattered around without any clustering. Hence, Figure 3 shows only the scores plot built using PC1 *vs*. PC2 and PC2 *vs*. PC3 by the raw and five treated counterparts.

As described above, each of the five locations was represented by two soil samples of brown and yellowish-brown, respectively. Hence, the most desired scores plot shall present five clusters by the location of origin. Unfortunately, none of the plot shown in Figure 3 produced five but three clusters.

Yellowish-brown BL soil was always away from the remaining samples, including the brown BL soil. This is expected since it possessed a highly individualistic chromatographic profile (check Supplementary Figure S2). However, it was observed to cluster with the majority of samples when inspected through a PC3 *vs*. PC2 scores plot. In other words, yellowish-brown BL soil still shares similarities with other soils in the RT window dominating the PC2. Although the yellowish-brown BL soil was consistently separated from the rest regardless of the type of sub-datasets, it is worth noting that its intra-sample variation was improved after being processed by the five BC algorithms.

Table 1. List of baseline correction (BC) algorithms evaluated in this study.

Algorithm	Description	Parameters
AsLS	An iterative algorithm applying second derivative constraints	Second derivative constraint, lambda = 6 Weighting of positive residuals, $P = 0.05$ Maximum number of iterations = 20
FP	An iterative algorithm using suppression of baseline by means in local windows	Second derivative penalty for primary smoothing =3 Number of iterations in suppression loop =10 Half width of local windows = 20
iRLS	An algorithm with primary smoothing and repeated baseline suppressions and regressions with second derivative constraint	Second derivative constraint for primary smoothing = 5 Second derivative constraint for secondary smoothing = 9 Maximum number of iterations = 200 Weighting of positive residuals = 0.05
MW	An algorithm finding medians in local windows and smoothing with Gaussian weighting	Window half width for local medians = 300
MPF	Polynomial fitting with baseline suppression relative to original spectrum	Degree of polynomial = 4 Maximum number of iterations = 100

AsLS: asymmetric least squares; FP: fill peak; iRLS: iterative restricted least squares; MW: median window; MPF: modified polynomial fitting.

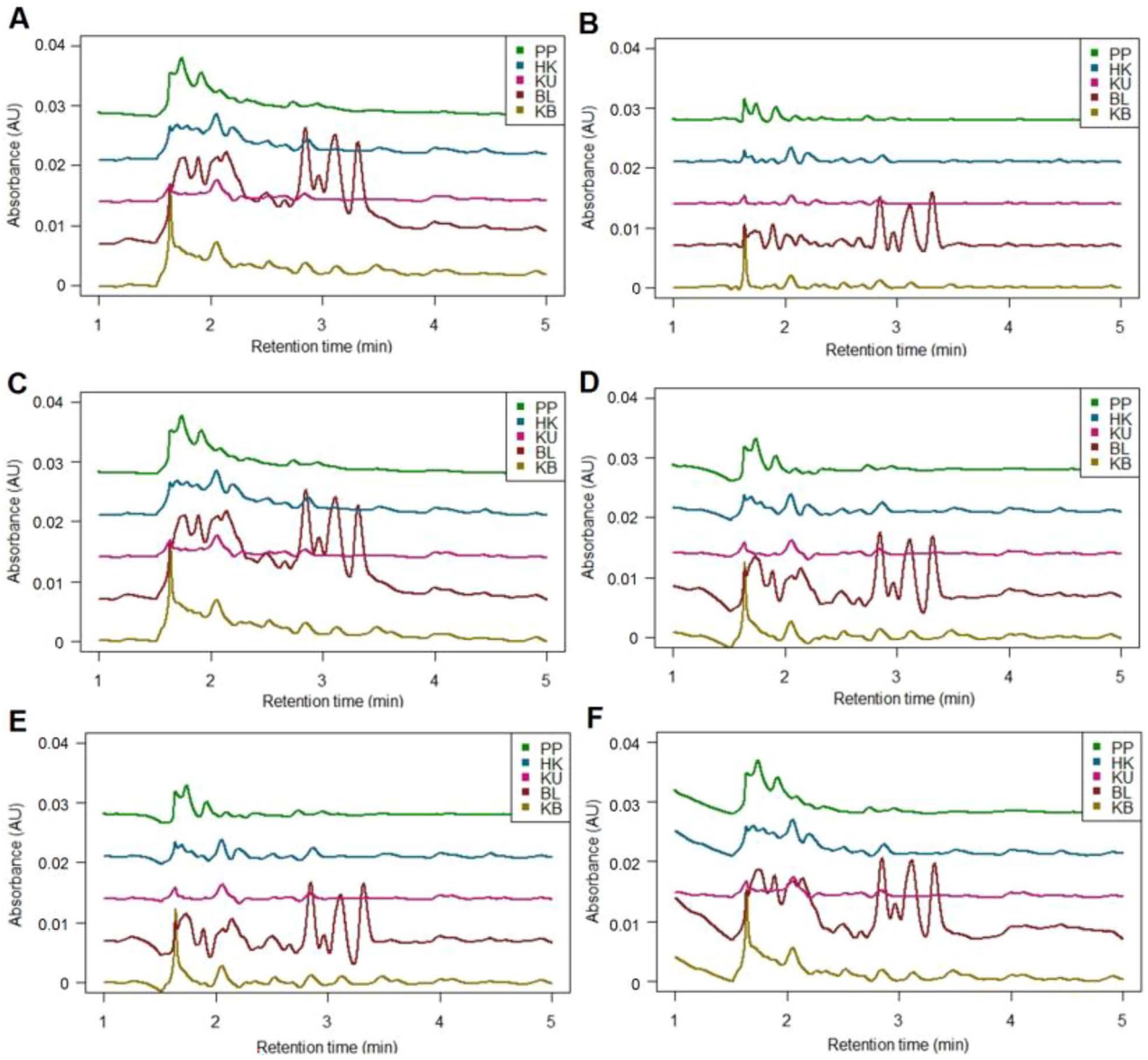

Figure 2 Mean chromatograms of raw (A) and that treated by AsLS (B), fill peak (C), iRLS (D), MW (E) and MPF (F) of sub-RT window data. KB: abandoned land nearby the Bangi commuter station (2.9008074"N, 101.7850107"E); BL: illegal trash dumping site (2.9015417"N, 101.7769922"E); KU: abandoned land nearby the Universiti Kebangsaan Malaysia (UKM) commuter station (2.9373368"N, 101.7907547"E); HK: UKM forest (2.9135556"N, 101.788083"E); PP: Fernarium UKM (2.9232222"N, 101.782722"E).

Next, it was observed that brown HK soil clustered with different samples dependent on the BC algorithm. Based on PC1 of the raw data and FP as well as MPF treated counterparts, it grouped with the brown soils of PP and KB. However, it clustered with the other soils when referred to the AsLS, iRLS, and MW treated data. In fact, the three algorithms were shortlisted as outperforming the other two algorithms based on mean chromatograms.

Overall, the soils seemed to cluster according to the colour rather than the location of origin. For instance, brown soils of three of the five locations, i.e. BL, PP, and KB, are hardly clustered with the corresponding yellowish-brown soils. Whilst the HK soils showed inconsistent intra-location variation, and the BC algorithms governed the degree of the intra-location variation. In contrast, brown and yellowish-brown KU soils were always clustered together. The low intra-location variation of KU could be attributed to the minimal human activities. The site is a hidden abandoned area away from the main road. In brief, the inter-locations variations of the five groups have been swamped by the exaggerated intra-location variations of PP, BL, and KB.

PLS-DA predictive performances

Supplementary Table 1 summarizes the mean and standard deviation of predictive accuracy rate estimated by the training and testing samples based on the PLS-DA models constructed using the first five PLS components. Surprisingly, AsLS treated data scored the lowest accuracy rates in both the samples and the raw sub-dataset outperformed the four treated sub-datasets. Nonetheless, the raw sub-dataset exhibited the highest standard deviation value denoting unstable performances and a high risk of overfitting.

The underperformance of AsLS could be explained by the drastic changes of the overall chromatographic patterns.

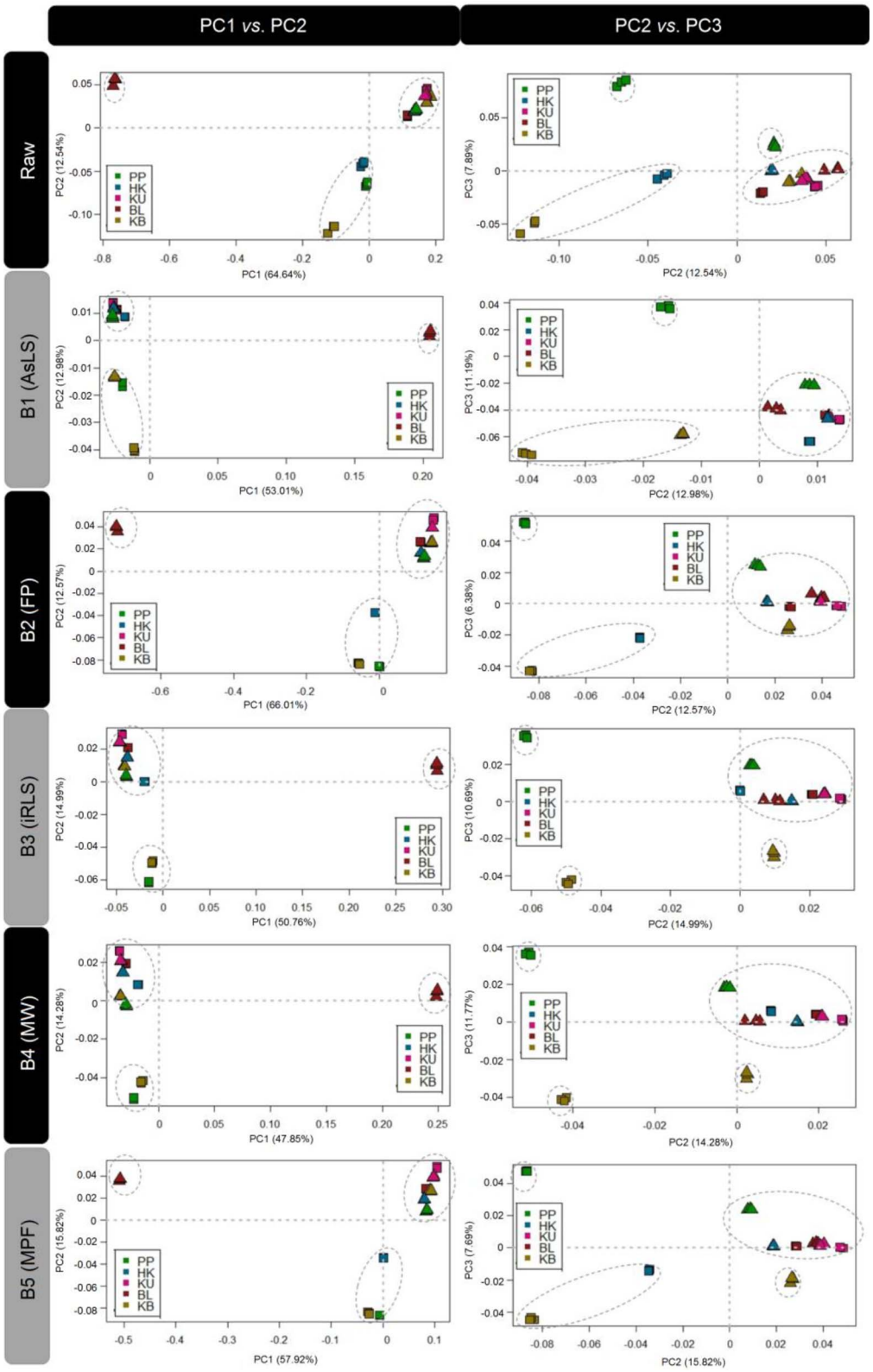

Figure 3 Scores plots of PCA showing spatial distribution of brown (squares) and yellowish-brown (triangles) soils presented by the raw and treated counterparts. KB: abandoned land nearby the Bangi commuter station (2.9008074"N, 101.7850107"E); BL: illegal trash dumping site (2.9015417"N, 101.7769922"E); KU: abandoned land nearby the Universiti Kebangsaan Malaysia (UKM) commuter station (2.9373368"N, 101.7907547"E); HK: UKM forest (2.9135556"N, 101.788083"E); PP: Fernarium UKM (2.9232222"N, 101.782722"E).

In other words, the algorithm has introduced artefacts that potentially degraded the performance of the original data. This alarmed the users to not rely upon only the visual exploratory tool in assessing the performance of a BC algorithm and data preprocessing method in general.

In order to gain more insight into the merits of BC algorithms in the UPLC data of soil, the modelling was repeated with another 1 000 pairs of training and testing samples by considering only the top three BC algorithms, i.e. FP, MW, and MPF, together with the raw counterpart. The mean and standard deviation of predictive accuracy rate shown in Table 2 are presented by four models constructed incrementally considering the first five PLS components. Moreover, the impact of second derivative penalty for primary smoothing in FP and window half-width for local medians in MW were carefully evaluated by considering varying values. The results indicated the parameters caused negligible effects (results not shown).

As expected, prediction accuracy is improved as more PLS components are considered, regardless of the input data. By referring to the prediction accuracy obtained using the training samples, one can see that MW tends to outperform the remaining sub-datasets except when modelled using the first five PLS components, where the raw counterpart emerged to be the outstanding sub-datasets. Nonetheless, MW also achieved over 0.9 accuracies. In contrast, MW consistently outperformed the remaining sub-datasets based on the prediction accuracy estimated using the testing set, which is also associated with the lowest standard deviation values. Hence, it is determined that MW is the excellent BC algorithm in this study.

Discussions

UPLC, one type of liquid chromatography, was employed in this work to obtain chemical profiles of 10 brownish soils from five proximity locations in Malaysia. Even though they presented different UPLC profiles allowing discrimination of soils by the location of origin, it is seen from Figure 2 that all the UPLC chromatograms showed varying fluctuations on the baselines. The baseline drift could deteriorate the discrimination of the soil samples *via* statistical predictive modelling.

According to Bos et al. [17], baseline drift in liquid chromatography data is mainly caused by small fluctuations in the flow rate and the mobile-phase composition; meanwhile, drifting of baseline in a gas chromatogram principally results from the variations in the flow rate and temperature-induced 'bleeding' of the column (i.e. stationary phase). Apparently, baseline problems can be of mechanical and chemical origins. Those arising from mechanical defects could be minimized or removed by technical measures, e.g. cleaning the flow cell in a UV detector [20]. However, chemical-originated defects are hard to be controlled but could be eliminated through a suitable BC algorithm. The BC algorithm typically aims to improve data performance by removing uneven amplitude shifts across the retention time of chromatograms [21].

Herein, five BC algorithms, i.e. AsLS, FP, iRLS, MW, and MPF, in discriminating the 30 UPLC chromatograms of brownish soils by five locations of origin were carefully assessed *via* visual inspection on mean chromatograms and scores plot of PCA as well as predictive modelling by PLS-DA method. On the other hand, the performances of the five BC algorithms in infrared spectral data of pen inks have been reported elsewhere [23]. The authors concluded that AsLS, FP, and MW algorithms outperformed iRLS and MPF algorithms. In contrast, this work found that MW outperformed the remaining algorithms, including the AsLS, which have often been reported as good in correcting infrared [24] and Raman [25] spectral data.

Mathematically, MW corrects the baseline by finding medians in local windows and smoothing with Gaussian weighting. The merits of MW are: (i) do not need to discriminate peak from noise; (ii) require no assumption regarding the source or functional form of the distortion [22]. On the other hand, AsLS is an iterative algorithm applying second derivative constraints in removing baseline artefacts. AsLS does not require peak finding because it sets the weights asymmetrically, updated iteratively. Technically, MW relies on only one parameter, i.e. window half-width for local medians, while AsLS has two parameters, i.e. second derivative constraint and weighting of positive residuals. Herein, we have not optimized AsLS based on the two parameters but employed the default values. Meanwhile, the only parameter of MW was assessed carefully using three different values since no default value is available. Hence, it seems like the underperformance of AsLS was not purely due to the unoptimized parameters but the inherent properties of the algorithm.

To date, the community of chromatographic users, especially UPLC data, seldom place considerable attention on

Table 2. Mean (standard deviation) predictive accuracy rate of partial least squares-discriminant analysis (PLS-DA) models estimated from 1 000 pairs of training and testing sets.

Algorithm	PLS1–2	PLS1–3	PLS1–4	PLS1–5
Training set				
Raw	0.4697 (0.0694)	0.6535 (0.0639)	0.8205 (0.0674)	**0.9255** (0.0573)
FP	0.4647 (0.0678)	0.7002 (0.0636)	0.8325 (0.0569)	0.9222 (0.0499)
MW	**0.5637 (0.0531)**	**0.7380 (0.0485)**	**0.8311 (0.0548)**	0.9088 (0.0590)
MPF	0.4667 (0.0662)	0.7045 (0.0616)	0.8315 (0.0562)	0.9175 (**0.0498**)
Testing set				
Raw	0.3061 (0.1430)	0.5214 (0.1668)	0.7071 (0.1828)	0.7889 (0.1768)
FP	0.3214 (0.1406)	0.5464 (0.1681)	0.7746 (0.1598)	0.7880 (0.1640)
MW	**0.3389 (0.1195)**	**0.6089 (0.1355)**	**0.7864 (0.1416)**	**0.7920 (0.1542)**
MPF	0.3208 (0.1438)	0.5620 (0.1678)	0.7805 (0.1546)	0.7893 (0.1571)

FP: fill peak; MW: median window; MPF: modified polynomial fitting. Bold value denotes the most desired PLS-DA model, i.e., highest accuracy rate and lowest standard deviation value.

optimizing the data *via* BC algorithms. This could be partly attributed to the conventional practice of deploying peak table data rather than pixel-level data for interpretation. By modelling the data composing only peaks selected carefully and manually by the researchers, interferences caused by uneven baseline can be detrimental. However, Riquelme et al. [8] have demonstrated that an automatic analysis pipeline is more feasible with pixel-level rather than peak table data. Hence, this study has signified the role of BC algorithms in processing pixel-level UPLC data.

Undeniably, the positive impact caused by the outstanding BC algorithm, i.e. MW, seen in this work is detrimental. This could be because the UPLC data studied herein is relatively small. Hence, to elucidate the role of BC algorithms in pixel-level data more clearly, future work shall attempt to evaluate the performance of the three best-performing BC algorithms with bigger UPLC data. Next, it is crucial to emphasize that most BC algorithms have multiple parameter values worth optimizing.

Conclusion

The five BC algorithms performed differently in correcting baseline drift of the UPLC data of soils. AsLS has improved the baseline of the raw chromatograms optimally. However, PLS-DA modelling denoted that MW is more outstanding than the AsLS sub-dataset. In conclusion, MW is the most desired option for the studied data.

Acknowledgements

The authors like to acknowledge the contributions of Anas, Ameeta, and Syahiera for performing the soil sampling, extraction, and data collection.

Authors' contributions

Muhamad Adibbin Ahmad drafted the manuscript. Nur Ain Najihah Mohd Rosdi performed the statistical analysis. Nadirah Binti Abd Hamid and Ab Aziz Ishak performed UPLC analysis. Loong Chuen Lee supervised the research. Hukil Sino conceived the experiment and the initial experimental design. Muhamad Adibbin Ahmad, Nadirah Binti Abd Hamid, and Nur Ain Najihah Mohd Rosdi were undergraduate and post-graduate students, registered under the forensic science program at UKM and supervised by Loong Chuen Lee and Hukil Sino.

Compliance with ethical standards

Ethical approval was gained from the UKM CRIM.

Disclosure statement

None declared.

Funding

The work was supported by UKM CRIM under Grant [GUP-2020-085].

References

1. Pye K. Geological and soil evidence: forensic applications. Boca Raton (FL): CRC Press, 2007.
2. Fitzpatrick RW. Soil: forensic analysis. In: Jamieson A, Moenssens A, editors. Wiley encyclopedia of forensic science. Chichester (UK): Wiley, 2009. p. 2377–2388.
3. Salih C, Ali CK, Ismail C, et al. SEM-EDS analysis and discrimination of forensic soil. Forensic Sci Int. 2004;141:33–37.
4. Sangwan P, Nain T, Singal K, et al. Soil as a tool of revelation in forensic science: a review. Anal Methods. 2020;12:5150–5159.
5. Xu X, Du C, Ma F, et al. Forensic soil analysis using laser-induced breakdown spectroscopy (LIBS) and Fourier transform infrared total attenuated reflectance spectroscopy (FTIR-ATR): principles and case studies. Forensic Sci Int. 2020;310:110222.
6. Profumo A, Agnese G, Sonia AG, et al. GC-MS qualitative analysis of the volatile, semivolatile and volatizable fractions of soil evidence for forensic application: a chemical fingerprinting. Talanta. 2020;219:121304.
7. McCulloch G, Dawson LA, Ros JM, et al. The discrimination of geoforensic trace material from close proximity locations by organic profiling using HPLC and plant wax marker analysis by GC. Forensic Sci Int. 2018;288:310–326.
8. Riquelme G, Zabalegui N, Marchi P, et al. A Python-based pipeline for preprocessing LC-MS data for untargeted metabolomics workflows. Metabolites. 2020;10:416.
9. Lee LC, Liong C-Y, Jemain AA. A contemporary review on Data Preprocessing (DP) practice strategy in ATR-FTIR spectrum. Chemom Int Lab Syst. 2017;163:64–75.
10. Md Ghazi MG, Lee LC, Sino H, et al. Review of contemporary chemometric strategies applied on preparing GC-MS data in forensic analysis. Microchem J. 2022;181:107732.
11. Slosse A, Van Durme F, Samyn N, et al. Evaluation of data preprocessing for the comparison of GC-MS chemical profiles of seized cannabis samples. Forensic Sci Int. 2020;310:110228.
12. Aloglu AK, Peter BH, Saliha S, et al. Chemical profiling of floral and chestnut honey using high-performance liquid chromatography-ultraviolet detection. J Food Compost Anal. 2017;62:205–210.
13. Ameeta NE. Pembeza Layan Sampel Tanah Keperangan Dengan Menggunakan Teknik Kromatografi Cecair Berprestasi Ultra (UPLC). [Distinguishing soil samples with reddish tones using ultra-performance liquid chromatography (UPLC)] BSc Thesis. Malaysia: Universiti Kebangsaan Malaysia, 2020. Malaysia.
14. Anas Z. Pembeza Layan Sampel Tanah Kemerahan Dengan Menggunakan Teknik Kromatografi Cecair Berprestasi Ultra (UPLC). [Distinguishing soil samples with reddish tones using ultra-performance liquid chromatography (UPLC)] BSc Thesis. Malaysia: Universiti Kebangsaan Malaysia, 2020.
15. Syahiera K. Pembeza Layan Sampel Tanah Perang Kekuningan Dengan Menggunakan Teknik Kromatografi Cecair Berprestasi Ultra (UPLC). [Differentiating soil samples with yellowish brown tones using ultra-performance liquid chromatography (UPLC)] BSc Thesis. Malaysia: Universiti Kebangsaan Malaysia, 2020. Malaysia.
16. Lee LC, Ishak AA, Nai Eyan A, et al. Forensic profiling of non-volatile organic compounds in soil using ultra-performance liquid chromatography: a pilot study. Forensic Sci Res. 2022;7:761–773.
17. Bos TS, Knol WC, Molennar SRA, et al. Recent applications of chemometrics in one- and two-dimensional chromatography. Separation Sci. 2020;43:1678–1727.
18. R Core Team. (2019). R: A Language and Environment for Statistical Computing. R Version 3.6.2 (12 December 2019). R Foundation for Statistical Computing: Vienna. Austria. Available from: https://www.Rproject.org/.
19. Lee LC, Liong C-Y, Jemain AA. Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps. Analyst. 2018;143:3526–3539.

20. Agilent Technologies. Eliminating Baseline Problems. 2007. Available from: https://www.agilent.com/cs/library/Support/Documents/Baseline_problems.pdf
21. Liland KH, Mevik B-H, Canteri R. Baseline correction of spectra. R Package 'baseline', Version 1.3-1, 2020. Available from: https://cran.r-project.org/web/packages/baseline/baseline.pdf
22. Friedrichs MS. A model-free algorithm for the removal of baseline artifacts. J Biomolecular NMR. 1995;5:147–153.
23. Lee LC, Liong C-Y, Khairul O, et al. Effects of baseline correction algorithms on forensic classification of paper based on ATR-FTIR spectrum and principal component analysis (PCA). Pertanika J Sci Tech. 2017;182:767–774.
24. Lee LC, Liong C-Y, Jemain AA. Effects of data preprocessing methods on classification of ATR-FTIR spectra of pen inks using partial least squares-discriminant analysis (PLS DA). Chemom Intel Lab Syst. 2018;182:90–100.
25. Korepanov VI. Asymmetric least-squares baseline algorithm with peak screening for automatic processing of the Raman spectra. J Raman Spectrosc. 2020;51:2061–2065.