Supplementary Information

A Chemometric Method to Quickly and Efficiently Guarantee the Authenticity of Commercialized Cigarettes in Brazil

Lucas S. Rodrigues, ****,a,b*** ***Carlos G. Massone*** ***a*** ***and José Marcus de O. Godoy*** ***a***

[a]*Departamento de Química, Pontifícia Universidade Católica (PUC-Rio), Rua Marquês de São Vicente, 225, Gávea, 22451-900 Rio de Janeiro-RJ, Brazil*

[b]*Instituto Federal de Educação, Ciência e Tecnologia do Mato Grosso (IFMT-Campus Confresa) Avenida Vilmar Fernandes, 300, Jardim Éden, 78652-000 Confresa-MT, Brazil*

Chromatographic alignment procedure details

Figure S1 illustrates the histogram obtained for the investigated cigarette samples.

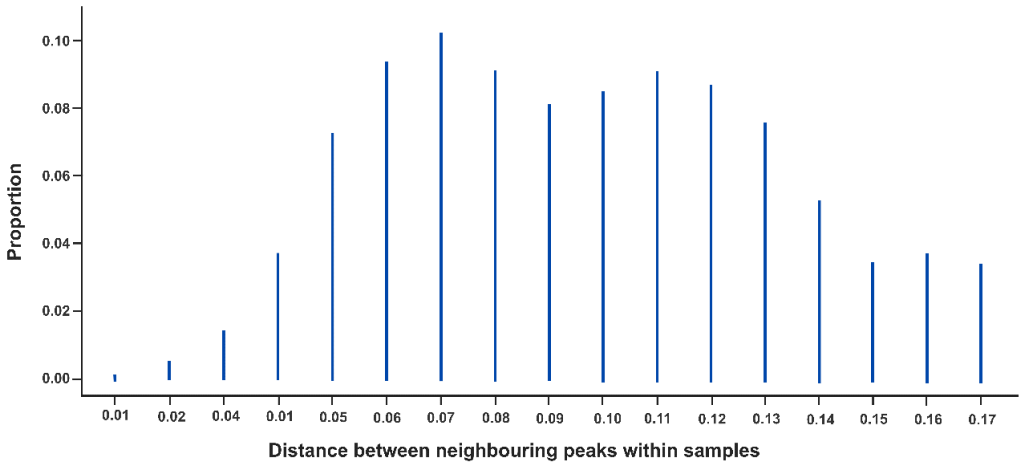

Figure S1. Histogram of retention-time (temporal) distances between chromatographic peaks measured across the cigarette tobacco samples (n = 96) analyzed by ultrasound-assisted extraction followed by GC-FID (UAE-GC-FID). The distribution is used to evaluate chromatographic reproducibility and to inform alignment parameter selection (retention-time units: insert units, e.g., seconds).

*e-mail: lucassrodrigues31.12.2014@gmail.com

A 0.07 min period was noted as the most frequent temporal distance between peaks, although temporal distances between 0.08 and 0.12 min were also verified relevant. Thus, 0.11 min was chosen for the subsequent max_linear_shift algorithm line analyses, to represent the greatest temporal distance with the greatest ratio between samples, aiding in t_R alignment effectiveness. Furthermore, few samples presented differences in $t_R \leq 0.05$ min. This parameter was thus adopted as the max_diff_peak2peak value. The package developers recommend that the allowed min_diff_peak2peak be twice the maximum distance, which was adopted herein. The rt_cutoff_low was set by excluding $t_R \leq 10$ min, as the solvent t_R was of ca. 6 min and no high intensity peaks were observed between 6 min < t_R < 10 min.

No samples were selected as reference for the alignment procedure, so the algorithm was allowed to select the sample with the greatest similarity to the others, set as the reference (reference = NULL). An analytical blank was used to fill the blanks argument. In addition, compounds present in only a single cigarette sample were eliminated from the data set using the delete_single_peak = TRUE argument, as they were not informative concerning similarity patterns.

A total of 275 chromatographic peaks were obtained, with 82 contained in the analytical blank, due to the fact that the model considers very small area values, close to noise. Another seven peaks were present in only one cigarette brand, with 186 peaks remaining following the chromatographic alignment procedure. Figure S2 displays the alignment procedure results for the proposed cigarette analysis method.

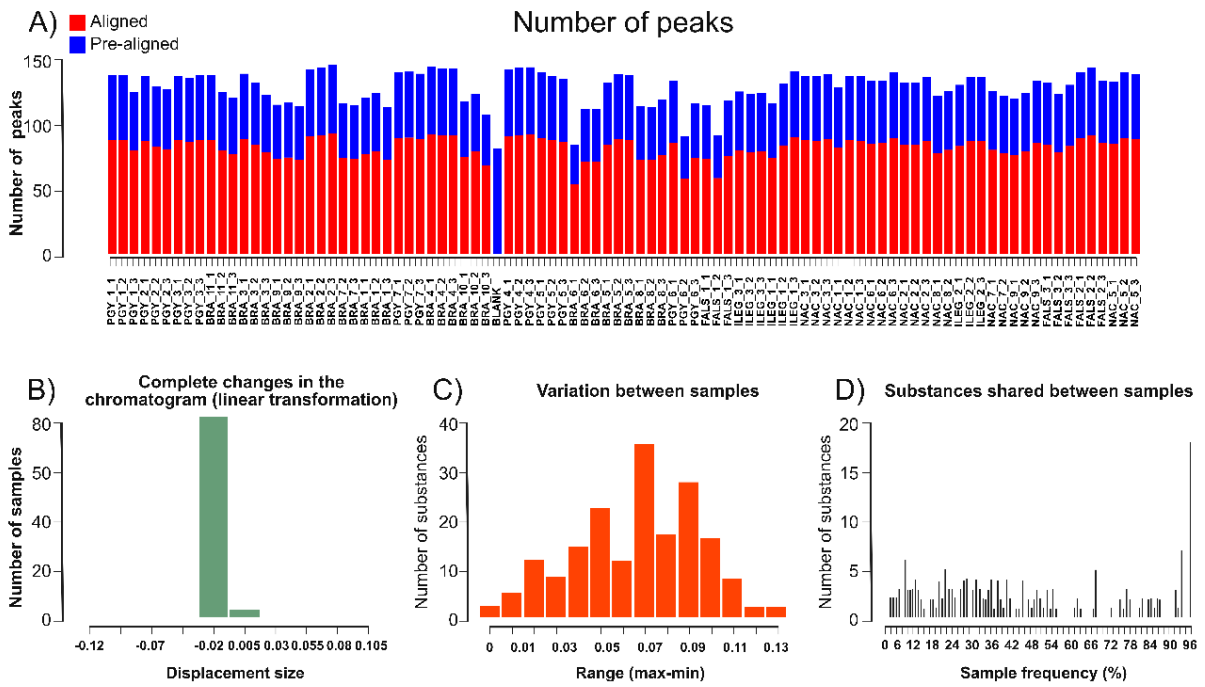

Figure S2. Diagnostic plots for retention-time alignment of chemical substances in the cigarette samples analyzed by UAE-GC-FID (n = total number of samples). Panels show: (a) number of peaks detected in each sample before and after alignment (assesses peak detection consistency); (b) histogram of linear retention-time shifts applied during alignment (shows magnitude and direction of corrections); (c) variation in retention times across samples for selected marker compounds (e.g., boxplots or line plots showing residual dispersion after alignment); and (d) frequency distribution of substances shared between samples (presence counts *per* compound), highlighting common *versus* rare compounds used in downstream clustering and classification.

Initially, as depicted in Figure S2a, each cigarette brand replicate contained about 150 detected compounds. After the alignment procedure, about 100 compounds remained. As noted in Figure S2b, most of the samples presented a linear shift very close to 0, which may indicate that no linear changes were necessary or that the transformations implemented herein were very small compared to the allowed range (0.11 min), suggesting this as an adequate linear shift value. Figure 2c indicates that most peak shifts of homologous substances range around 0.05 to 0.09 min, suggesting less variations than the chosen linear shift value of 0.11 min. This, in turn, demonstrates that the t_R of supposedly homologous peaks in the aligned data set shifted below the maximum limit, proving the effectiveness of the applied alignment procedure.

Another informative data, displayed in Figure S2d, is that only 6% of the samples contain only five compounds in common, with all other samples having at least 20 compounds in common. This proves that the method was able to extract a high number of common compounds among the investigated cigarette samples.

Figure S3 shows a heat map graph used to assess potential associations between the identified substances and the investigated cigarette samples.

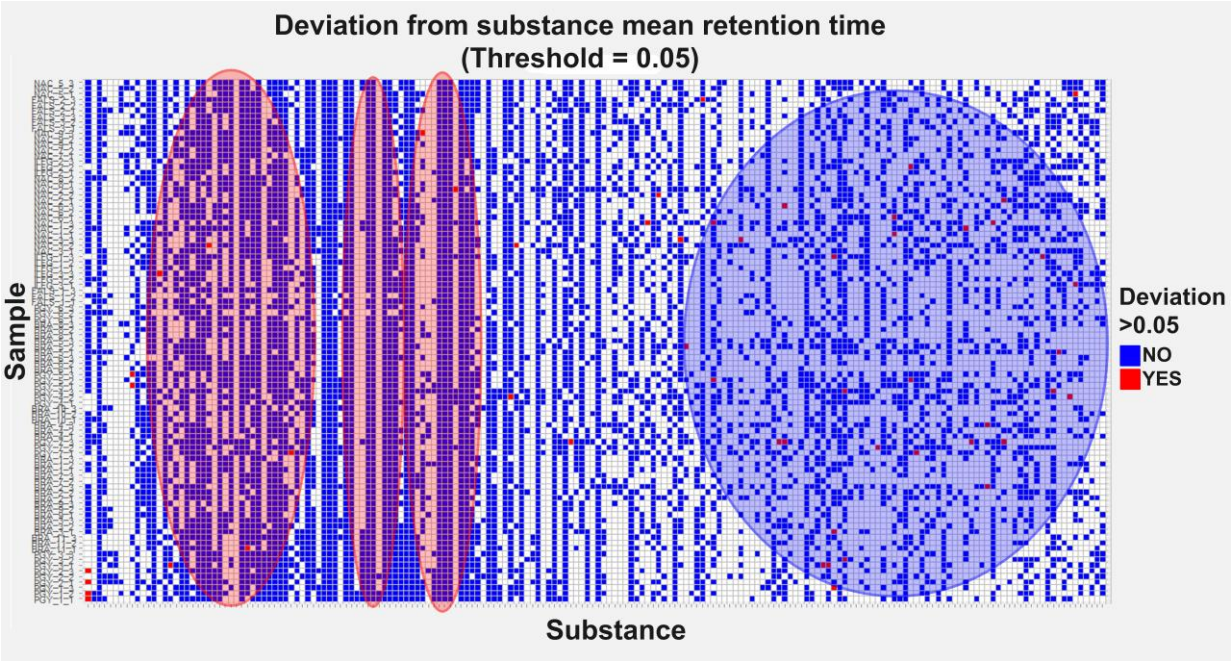

Figure S3. Heatmap of cigarette tobacco samples (n = 96), built from peak area measurements obtained by UAE-GC-FID. Rows correspond to detected chemical features and columns to individual samples (with top color bars indicating brand/group). Data were preprocessed by insert (e.g., log transformation and *per*-feature normalization) The heatmap highlights chemical patterns that assist discrimination cigarettes brands.

In this type of graph, the abscissa axis represents the different types of t_R detected after the alignment procedure and the ordinate axis represents the cigarette samples. It is classified as a presence and absence graph, where the blue filling indicates the presence of a certain compound. The red filling indicates that a certain t_R was not fully aligned by the max_linear_shift function, considered an undetected t_R outlier.

The heatmap indicates high data variability. The depicted red areas contain some white regions, meaning that all samples present the same t_R in this region, with few differences between the investigated cigarettes. However, a significant variability in the blue area is noted, with alternating present and absent compounds, aiding in differentiating cigarette groups.

Chromatographic peak alignment algorithm

```
library("GCalignR")

data_1<-read_peak_list ('ultrasound_extracts.txt', sep = "\t", rt_col_name='RT', check = T)

check_input (data = data_1, plot = T)

peak_interspace (data = data_1, rt_col_name = "RT",
            quantile_range = c (0, 0.8), quantiles = 0.05)

peak_data_aligned <- align_chromatograms (data = data_1,
                                  rt_col_name = "RT",
                                  rt_cutoff_low = 10,
                                  max_linear_shift = 0.11,
                                  max_diff_peak2mean = 0.05,
                                  min_diff_peak2peak = 0.1,
                                  blanks = "Blank_1",
                                  delete_single_peak = FALSE,
                                  write_output = NULL)

print(peak_data_aligned)

plot(peak_data_aligned)

gc_heatmap(peak_data_aligned)

scent<- norm_peaks (peak_data_aligned, conc_col_name = "Area", rt_col_name = "RT", out = "data. frame")

scent2 <- log (scent + 1)
```

Multicollinearity test

```
library(corrplot)[1]
correl<-cor (scent2 [,1:124])
correl2<-sapply (as.data. frame(correl), as. numeric)
correl2<-as.data. frame(abs(correl2))
str(correl2)

for (i in 1:124) {
  for (j in 1:124) {
```

```
    correl [i, j] <-ifelse (correl [i, j]>0.5, 1, 0)
  }
}
barplot (apply (correl, 2, sum))
apply (correl, 2, sum)
listagem<-data. frame (which (apply (correl, 2, sum)>=2))
lista<-rownames(listagem)
library(dplyr)[2]
scentnew<-scent2[,!(names(scent2)%in%lista)]
```

Cluster analysis setup

```
plot(hclust(dist(scale(scentnew), method = "ward. D2"))
```

LDA setup

```
library (MASS)[3]
library(lares)[4]

groups<-left(rownames(scentnew), 2)

scent2$groups<-
c("Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Multinati
onal","Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Multinational",
"Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Mult
inational","Multinational","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paragua
y","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Multinational","Multinational","
Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Parag
uay","Paraguay","Paraguay","Counterfeit","Counterfeit","Counterfeit","No_Health_Registration","No_Health_Registra
tion","No_Health_Registration","No_Health_Registration","No_Health_Registration","No_Health_Registration","Regi
onal","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Region
al","Regional","Regional","Regional","No_Health_Registration","No_Health_Registration","No_Health_Registration",
"Regional","Regional","Regional","Regional","Regional","Counterfeit","Counterfeit","Counterfeit","Counterfeit","Cou
nterfeit","Counterfeit","Regional","Regional","Regional")

groups_2<-
c("Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Multinati
onal","Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Multinational",
"Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Mult
inational","Multinational","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paragua
y","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Paraguay","Multinational","Multinational","
```

Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Multinational","Paraguay","Paraguay","Paraguay","Counterfeit","Counterfeit","Counterfeit","No_Health_Registration","No_Health_Registration","No_Health_Registration","No_Health_Registration","No_Health_Registration","No_Health_Registration","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Regional","Regional","No_Health_Registration","No_Health_Registration","No_Health_Registration","Regional","Regional","Regional","Regional","Regional","Counterfeit","Counterfeit","Counterfeit","Counterfeit","Counterfeit","Counterfeit","Regional","Regional","Regional")

Where: Regional = Small-Scale Regional Brands, Counterfeit = Cigarette brands that are counterfeits of Paraguayan brands, No Health Registration = Brands that have lost their ANVISA health registration, Multinational = Cigarette brands from large multinational cigarette companies, Paraguay = Brands smuggled from Paraguay.

LDA plotting

```
LLLnew<-lda (scentnew [,1:85], scent$grupos)
lda_predictionnew <- predict (LLLnew)
scalasnew<-as.data. frame(lda_predictionnew$x)
theme_set (theme_bw (base_size = 14))
ggplot (scalasnew, aes (x=LD1, y=LD2, colour=as. factor(groups_2))) +
  geom_point () +
  geom_text_repel(aes(label=groups_2)) +
  scale_colour_manual ("Origin", values=c ("navy", 'firebrick3', "red", "black", "green")) +
  xlab ("LD1 (61.07%)") +
  ylab ('LD2 (19.69%)') +
  theme (legend. position = c (0.9, 0.1), legend. margin = margin (5,5,5,5),
       legend. background = element_rect (fill='gray90', colour='black'))
cml<-table (scent$grupos, lda_prediction$class)
cml
```

Validation via Training-Testing Set Partitioning

```
ind <- sample (2, nrow(scentnew),
               replace = TRUE,
               prob = c (0.7, 0.3))

training <- scentnew [ind= =1,]
testing <- scentnew [ind= =2,]
groups_3<-left(rownames(training), 2)
```

The analyst will need to analyze the testing dataframe (exclusively in this step) to see which samples were randomly selected and classify them according to their initial prediction: testing$grupos<-c (samples classified according to the analyst's previous verification).

```
lda_model<-lda (training [,1:85], training$groups)
lda_pred <- predict (lda_model, testing$groups)
pred_classes <- lda_pred$class
conf_matrix <- table (testing$groups, pred_classes)
print(conf_matrix)
accuracy <- sum(diag(conf_matrix)) / sum(conf_matrix)
```

The sensitivity and specificity values of the chemometric model were calculated from the cml table for each class. The radar chart was constructed from these values.

Validation via ROC analysis and AUC evaluation

```
library(pROC)[6]
multiclass.roc (testing$groups, lda_model$posterior [,1], plot=TRUE, legacy. axes = TRUE, percent =TRUE, xlab="False Positive Percentage", ylab="True Positive Percentage", colorize = T)

multiclass.roc (testing$groups, lda_model$posterior [,2], plot=TRUE, legacy. axes = TRUE, percent =TRUE, xlab="False Positive Percentage", ylab="True Positive Percentage", colorize = T)

multiclass.roc (testing$groups, lda_model$posterior [,3], plot=TRUE, legacy. axes = TRUE, percent =TRUE, xlab="False Positive Percentage", ylab="True Positive Percentage", colorize = T)

multiclass.roc (testing$groups, lda_model$posterior [,4], plot=TRUE, legacy. axes = TRUE, percent =TRUE, xlab="False Positive Percentage", ylab="True Positive Percentage", colorize = T)

multiclass.roc (testing$groups, lda_model$posterior [,5], plot=TRUE, legacy. axes = TRUE, percent =TRUE, xlab="False Positive Percentage", ylab="True Positive Percentage", colorize = T)

multiclass.roc (testing$groups, lda_model$posterior [,6], plot=TRUE, legacy. axes = TRUE, percent =TRUE, xlab="False Positive Percentage", ylab="True Positive Percentage", colorize = T)
```

References

1. Wei, T.; Simko, V.; *R package 'corrplot': Visualization of a Correlation Matrix*, version 0.95; 2024. [Crossref]
2. Wickham, H.; François, R.; Henry, L.; Müller, K.; Vaughan, D.; *dplyr: A Grammar of Data Manipulation*; R package, version 1.1.4; 2025. [Crossref]
3. Venables, W. N.; Ripley, B. D.; *Modern Applied Statistics with S*, 4th ed.; Springer: NY, USA, 2002. [Crossref]
4. Lares, B.; *lares: Analytics & Machine Learning Sidekick. R package*, version 5.2.10. [Crossref]
5. Wickham, H.; François, R.; Henry, L.; Müller, K.; Vaughan, D/A; *dplyr: Uma Gramática de Manipulação de Dados*, Pacote R version 1.1.4; 2025. [Crossref]
6. Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.; Müller, M.; *BMC Bioinformatics* **2011**, *12*, 77. [Crossref]