Supplement of Atmos. Meas. Tech., 12, 2287–2312, 2019
https://doi.org/10.5194/amt-12-2287-2019-supplement

Atmospheric Measurement Techniques
Open Access
EGU

Supplement of

Analysis of functional groups in atmospheric aerosols by infrared spectroscopy: systematic intercomparison of calibration methods for US measurement network samples

Matteo Reggente et al.

Correspondence to: Satoshi Takahama (satoshi.takahama@epfl.ch)

Contents

S1 Atom apportionment matrix **2**

S2 Contributions to OM/OC **2**

S3 Relationship between PLS regression and Beer-Lambert absorption coefficients **3**

S4 Latent variables **3**

S5 Calibration curves for laboratory standards **5**

S6 PLS model selection **8**

S7 Comparisons of FG estimates **9**

S8 Variations in FG abundance with number of LVs used in the PLSbc model. **15**

S9 Variable Importance in Projection and Explained Variation in carbonyl calibration models **17**

S10 OM/OC probability distributions **18**

S11 Christiansen peak effect in anomalous clusters 19 and 20 **19**

S12 Variations of OC, FG and OM/OC with number of LVs for aCH with the PLSbc model **20**

S1 Atom apportionment matrix

The OC, OM, OM/OC ratio, and O/C ratio can be estimated by mapping functional group abundances to abundance of their atomic constituents. As discussed in Section 2, the moles of atoms are estimated from the moles of FG as $n_a = \sum_k \lambda_{ak} n_k$:

$$\begin{aligned} n_{\mathrm{C}} &= (0.5 \text{ or } 0)\, n_{\mathrm{aCOH}} + n_{\mathrm{cCOH}} + 0.5\, n_{\mathrm{aCH}} + n_{\mathrm{tCO}} \\ n_{\mathrm{O}} &= \qquad\qquad n_{\mathrm{aCOH}} + n_{\mathrm{cCOH}} \qquad\qquad + n_{\mathrm{tCO}} \\ n_{\mathrm{H}} &= \qquad\qquad n_{\mathrm{aCOH}} + n_{\mathrm{cCOH}} + \quad n_{\mathrm{aCH}} \end{aligned}$$

Values less than unity for carbon presumes polyfunctionality, and is intended to prevent multiple counting. The value by Russell and co-workers and (**?**) for $\lambda_{\mathrm{C,aCOH}}$ differ by 0.5. A value of 0.5 corresponds to the assumption that the carbon shares an aCOH bond with a single aCH bond, whereas a value of 0 corresponds to the assumption of a terminal saturated carbon in which it is accounted for by two aCH bonds. According to the analysis of **?**, we use a value of 0.5.

Carboxylic COH and carbonyl moeities characterized by individual bonds (**?**) can be related to their functional group equivalents (**?**) by:

$$\begin{pmatrix} n_{\mathrm{COOH}} \\ n_{\mathrm{naCO}} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} n_{\mathrm{cCOH}} \\ n_{\mathrm{tCO}} \end{pmatrix} . \tag{S1}$$

The contributions of COOH and naCO to the C, O, and H budget are as follows:

$$\begin{pmatrix} n_{\mathrm{C}} \\ n_{\mathrm{O}} \\ n_{\mathrm{H}} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 2 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} n_{\mathrm{COOH}} \\ n_{\mathrm{CO}} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} n_{\mathrm{cCOH}} \\ n_{\mathrm{tCO}} \end{pmatrix} .$$

S2 Contributions to OM/OC

Following notation used previously, n_k is the molar abundance of FG k, λ_{ak} is the number of atoms a associated with FG k, and m_a is the atomic mass of a. The OM/OC ratio can be defined as

$$\frac{OM}{OC} = \frac{\sum_k \sum_a m_a \lambda_{ak} n_k}{\sum_k m_{\mathrm{C}} \lambda_{\mathrm{C}\cdot k} n_k} .$$

For mathematical convenience, we will rearrange the expression above to:

$$\frac{OM}{OC} = 1 + \sum_k \zeta_k \left[\frac{n_k}{n_k + \sum_{k' \neq k} \left(\lambda_{\mathrm{C}\cdot k'} / \lambda_{\mathrm{C}\cdot k} \right) n_{k'}} \right] , \quad \text{where} \quad \zeta_k = \frac{\sum_a m_a \lambda_{ak}}{m_{\mathrm{C}} \lambda_{\mathrm{C}\cdot k}} .$$

We find that each term in the summation has the following limit (f is an arbitrary constant):

$$\lim_{n \to \infty} \zeta \left(\frac{n}{n+f} \right) \rightarrow \zeta$$

which suggests that as the abundance of any bond begins to dominate the composition, its contribution to the OM/OC ratio will approach a fixed value ζ. This limit explains the reason for the OM/OC contribution from aCH often approaches a value of $(1.008 \cdot 1)/(12.01 \cdot 0.5) = 0.168$, since its molar contribution is typically more than 60%.

S3 Relationship between PLS regression and Beer-Lambert absorption coefficients

Letting $n^{(a)}$ denote moles per unit area and ϵ denote molar absorption coefficient, the Beer-Lambert law is written as:

$$x_{ij} = \sum_k \epsilon_{jk} n_{ik}^{(a)} \tag{S2}$$

Consider two species:

- species c absorbing at wavenumber j
- species c' absorbing at both wavenumbers j and j'

The spectra defined as sum of absorbances is given as:

$$\begin{aligned} x_{ij} &= n_{ic}^{(a)} \epsilon_{jc} + n_{ic'}^{(a)} \epsilon_{jc'} \\ x_{ij'} &= n_{ic'}^{(a)} \epsilon_{jc'} \end{aligned} \tag{S3}$$

Letting f_{jc} denote the fraction of absorbance at a wavenumber j that is attributable to species c, we can define the areal density as:

$$\begin{aligned} n_{ic}^{(a)} &= x_{ij} f_{jc} \epsilon_{jc}^{-1} = \left(x_{ij} - n_{ic'}^{(a)} \epsilon_{jc'} \right) \epsilon_{jc}^{-1} = \left(x_{ij} - x_{ij'} \epsilon_{j'c'}^{-1} \epsilon_{jc'} \right) \epsilon_{jc}^{-1} \\ n_{ic'}^{(a)} &= x_{ij} f_{jc'} \epsilon_{jc'}^{-1} = x_{ij'} \epsilon_{j'c'}^{-1} \end{aligned} \tag{S4}$$

Then we can interpret coefficients from the multilinear regression model as a combination of absorbances. For species k:

$$\begin{aligned} n_{ic}^{(a)} &= x_{ij} b_{jc} + x_{ij'} b_{j'c} \\ b_{jc} &= \epsilon_{jc}^{-1} \\ b_{jc'} &= -\epsilon_{j'c'}^{-1} \epsilon_{jc'} \epsilon_{jc}^{-1} \end{aligned} \tag{S5}$$

As species c' is an interferent for species c, the regression coefficient is negative.

S4 Latent variables

We rewrite equation set in array notation:

$$\begin{aligned} \mathbf{X} &= \mathbf{T}\mathbf{P}^T + \mathbf{E}_X \\ \mathbf{y} &= \mathbf{T}\mathbf{q}^T + \mathbf{e}_y \ . \end{aligned}$$

For fitted data, the explained variation by component k is defined as:

$$\begin{aligned} EV_{X,k} &= \frac{\left(\mathbf{p}_k^T \mathbf{p}_k \right) \cdot \left(\mathbf{t}_k^T \mathbf{t}_k \right)}{\mathrm{Tr}\left(\mathbf{X}^T \mathbf{X} \right)} \\ EV_{y,k} &= \frac{q_k^2 \mathbf{t}_k^T \mathbf{t}_k}{\mathbf{y}^T \mathbf{y}} \end{aligned}$$

For the prediction set (denoted by subscript p), which are the ambient samples in this case, the loadings remain the same:

$$\mathbf{X}_p = \mathbf{T}_p\mathbf{P}^T + \mathbf{E}_{Xp}$$
$$\mathbf{y}_p = \mathbf{T}_p\mathbf{q}^T + \mathbf{e}_{yp} \ .$$

Note that $\mathbf{X}_p$ and $\mathbf{y}_p$ are centered by the means of $\mathbf{X}$ and $\mathbf{y}$, respectively. In addition to explained variance in $\mathbf{X}_p$, we use the sum-of-squares (SS) as an equivalent metric to understand the importance of each LV toward prediction of $\mathbf{y}_p$ (there is no reference for ambient prediction set samples):

$$EV_{Xp,k} = \frac{\left(\mathbf{p}_k^T\mathbf{p}_k\right)\cdot\left(\mathbf{t}_{p,k}^T\mathbf{t}_{p,k}\right)}{\mathrm{Tr}\left(\mathbf{X}_p^T\mathbf{X}_p\right)}$$
$$SS_{yp,k} = q_k^2\mathbf{t}_{p,k}^T\mathbf{t}_{p,k}$$

In PLSr, nearly 100% of the variation in laboratory standard spectra is explained by the selected LVs, though for ambient samples this value varies between 80–100%, except for tCO for which it is 65%. The value for tCO is lower because the narrow absorption region of tCO does not require much of the variation in the rest of the spectra to be modeled (a high EV_X is not a requirement for a good calibration model with PLS). The first latent variable typically explains >60% of the variation in both laboratory and ambient sample spectra except for tCO, where it models less than 50%. The reason for this high value is that this first LV models variability in the PTFE contribution to the spectral signal (illustrated in Figure Sx). Typically, five fewer LVs are required to reach the explained sum-of-squares for the response variable (FG) in laboratory standard spectra than ambient sample spectra because the increasing complexity of atmospheric samples. In the neighborhood of the selected solution, solutions have high correlations with the spectra but varies substantially in the slope (i.e., the variance in the predicted solution is manifested primarily in the slope).

S5 Calibration curves for laboratory standards

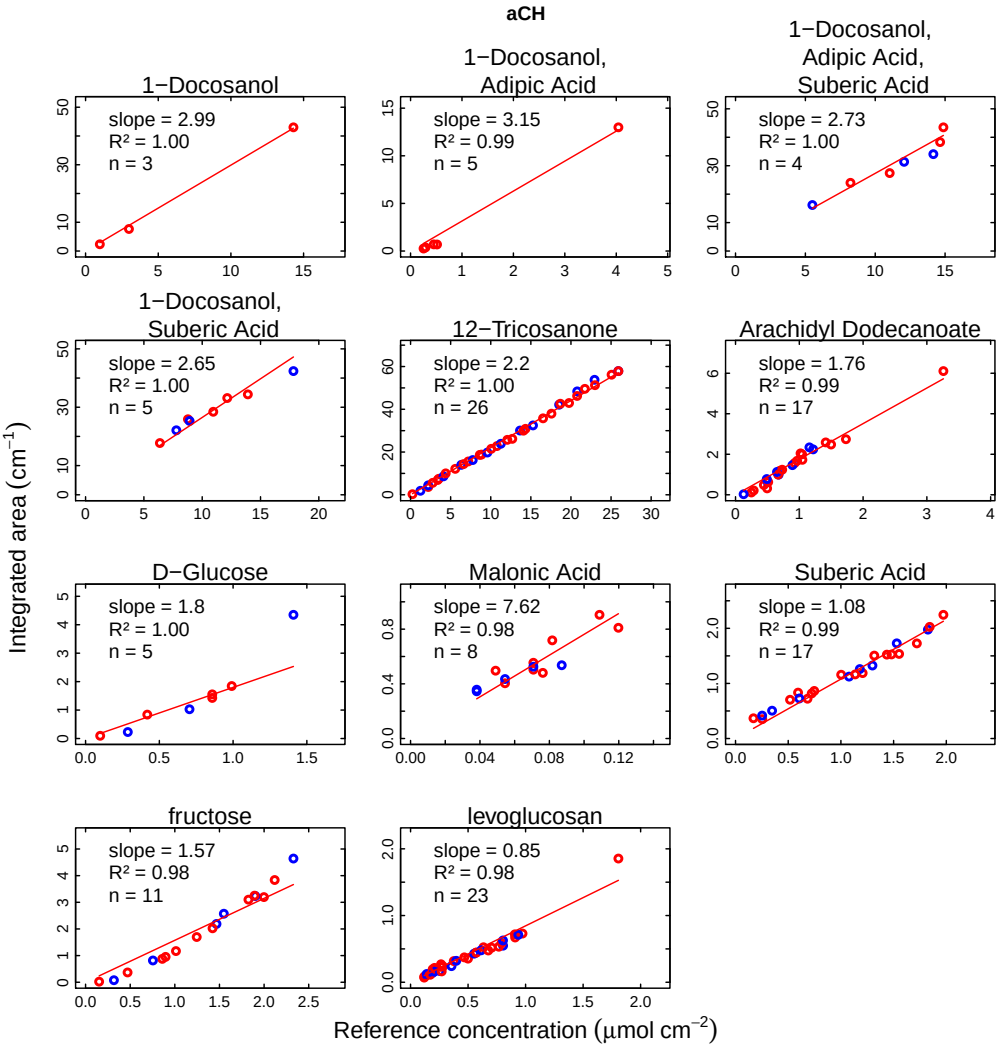

Figure S1: Laboratory standards calibration curves for aCH.

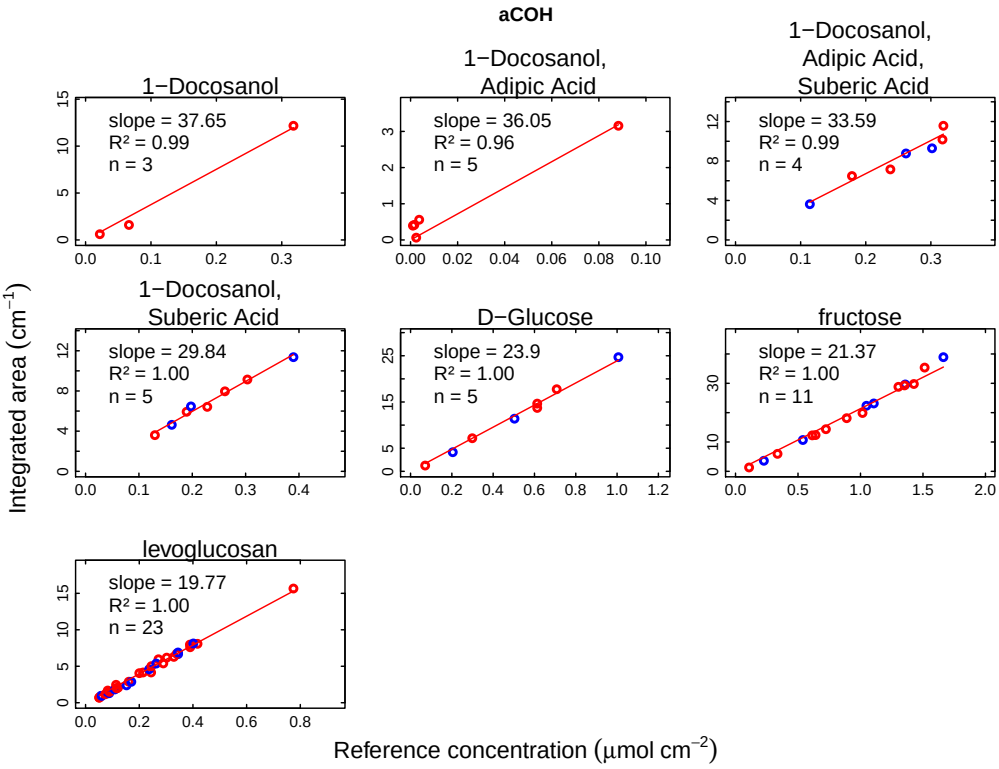

Figure S2: Laboratory standards calibration curves for aCOH.

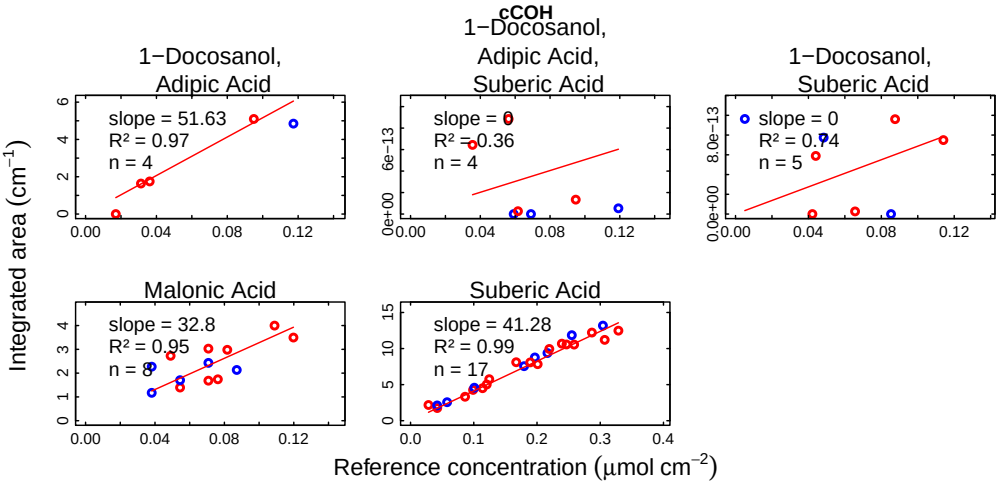

Figure S3: Laboratory standards calibration curves for cCOH.

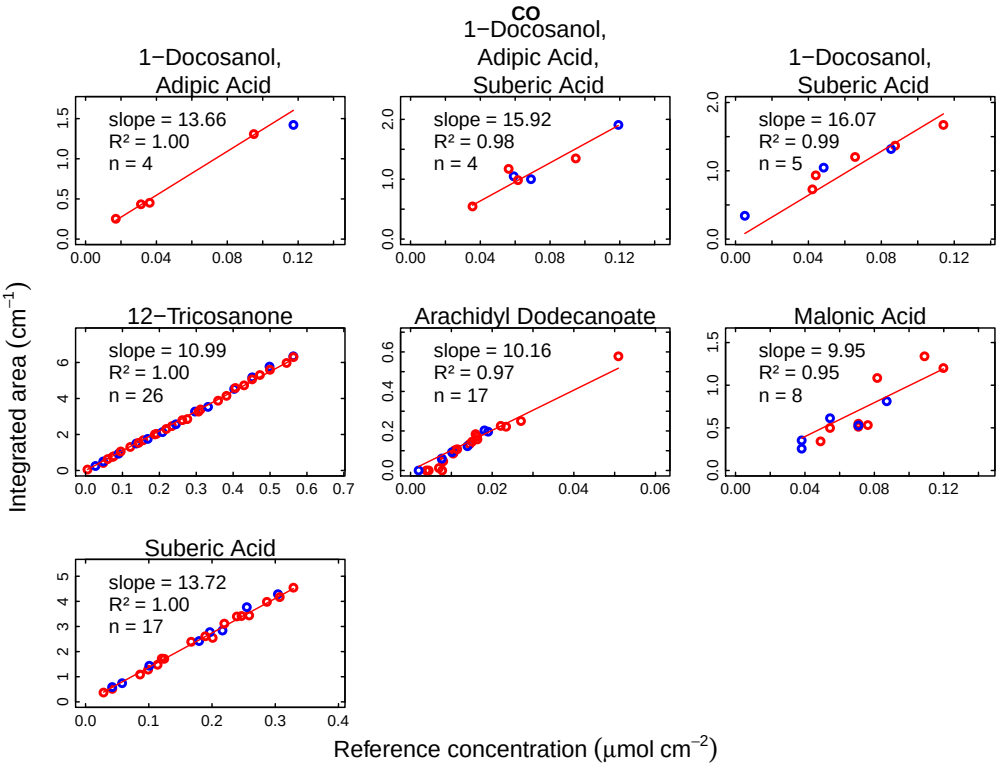

Figure S4: Laboratory standards calibration curves for CO.

S6 PLS model selection

Table S1: PLS models: selected number of latent variables (LVs) using the method proposed by **?**. In bracket the number of LVs selected by the minimum RMSE.

Model	Number of LVs						
	aCOH	cCOH	aCH	tCO	naCO	COOH	NH
PLSr	12 (30)	18 (18)	12 (46)	17 (18)	17 (21)	17 (18)	19 (35)
PLSbc	12 (16)	9 (10)	11 (34)	9 (10)	10 (12)	9 (10)	11 (17)

S7 Comparisons of FG estimates

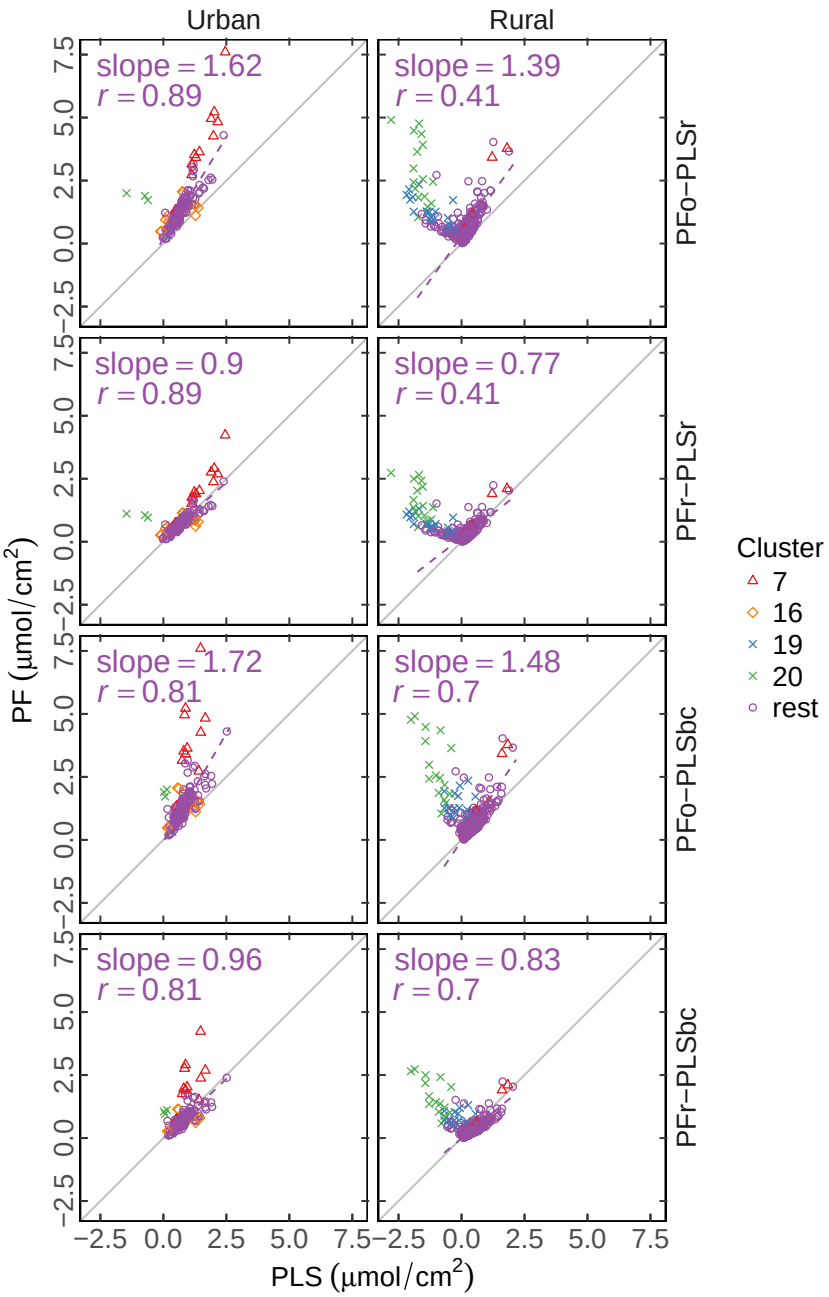

Figure S5: aCOH

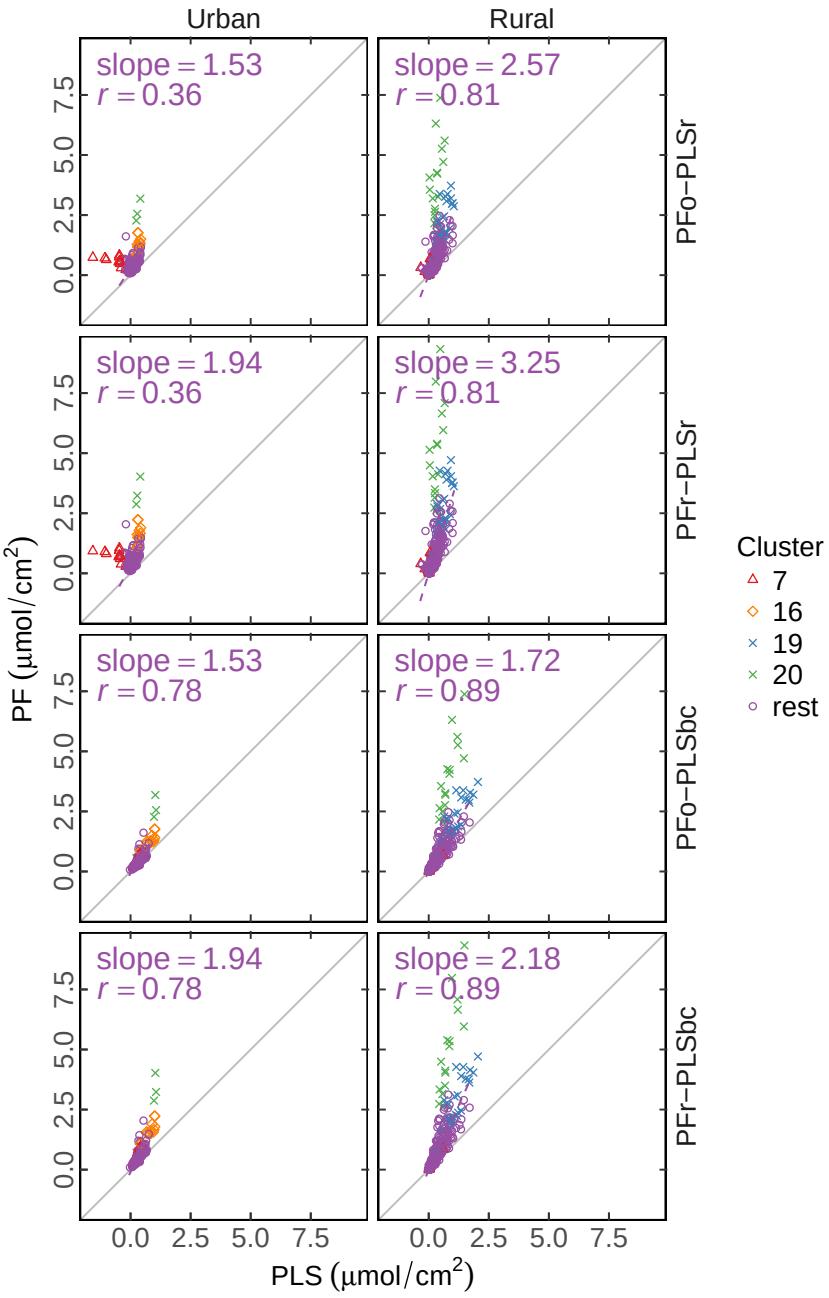

Figure S6: cCOH

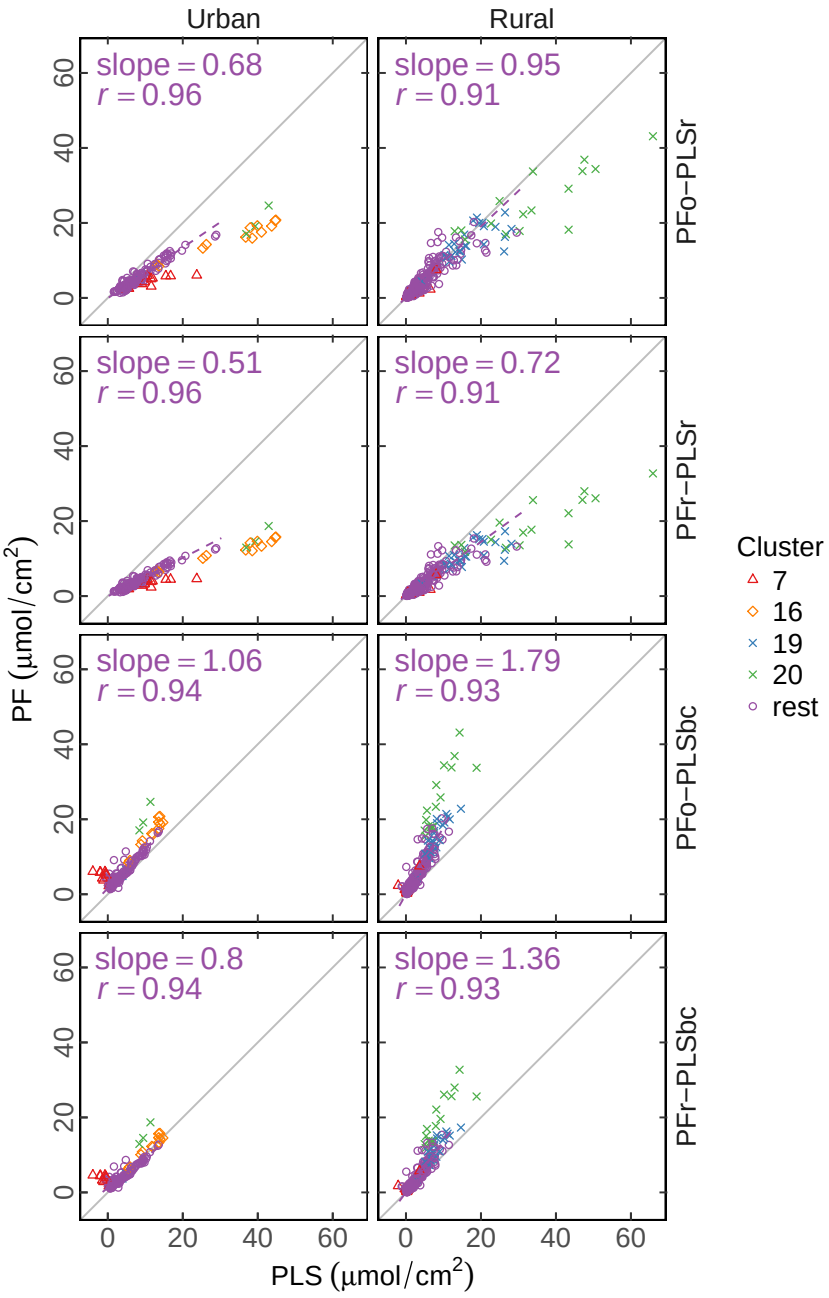

Figure S7: aCH

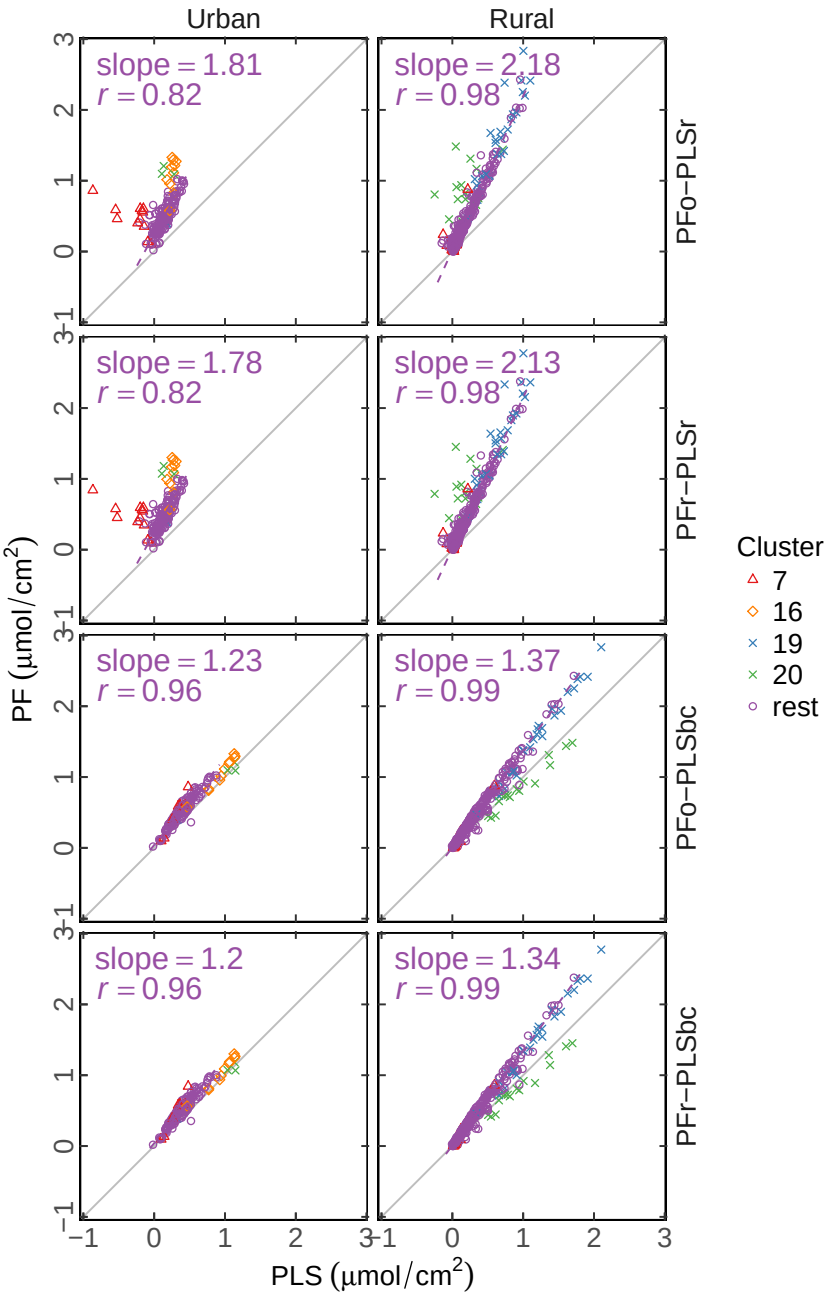

Figure S8: tCO

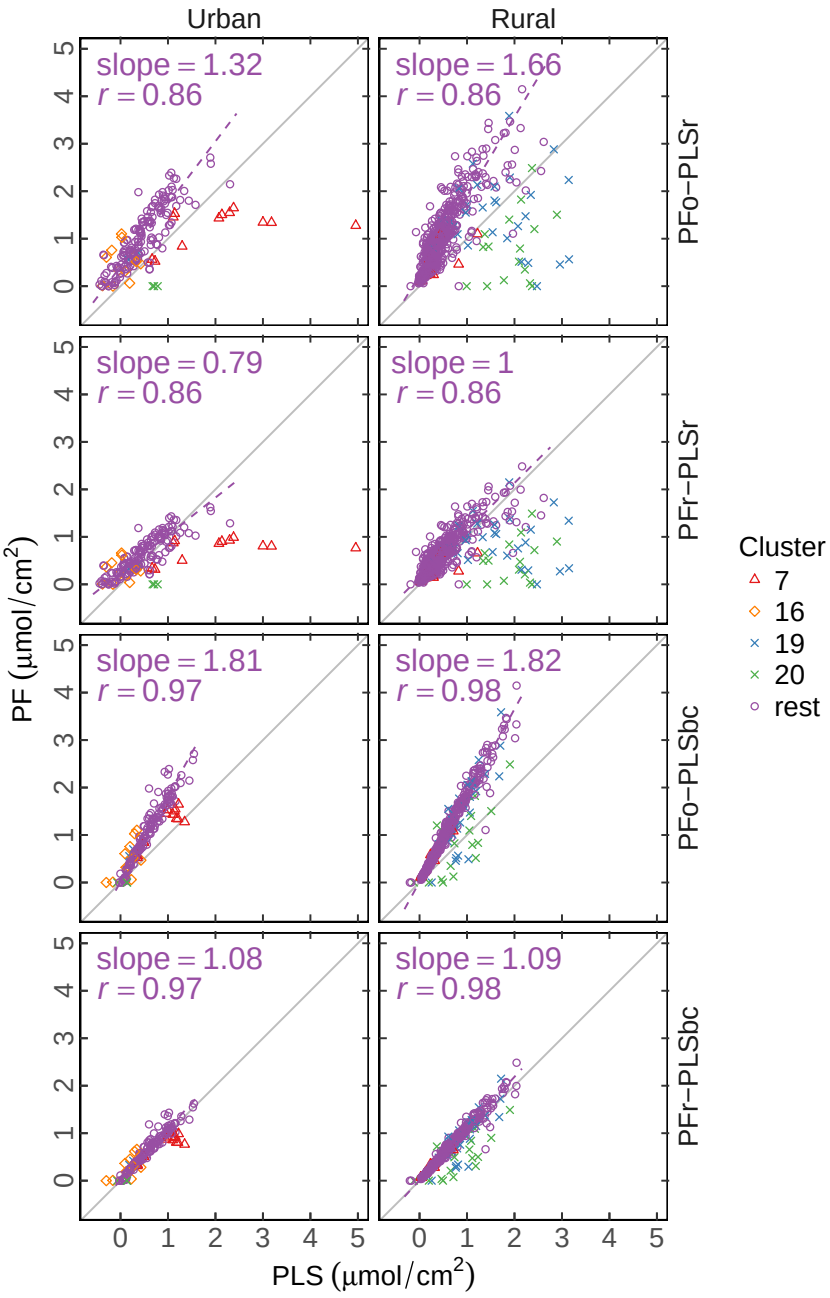

Figure S9: iNH

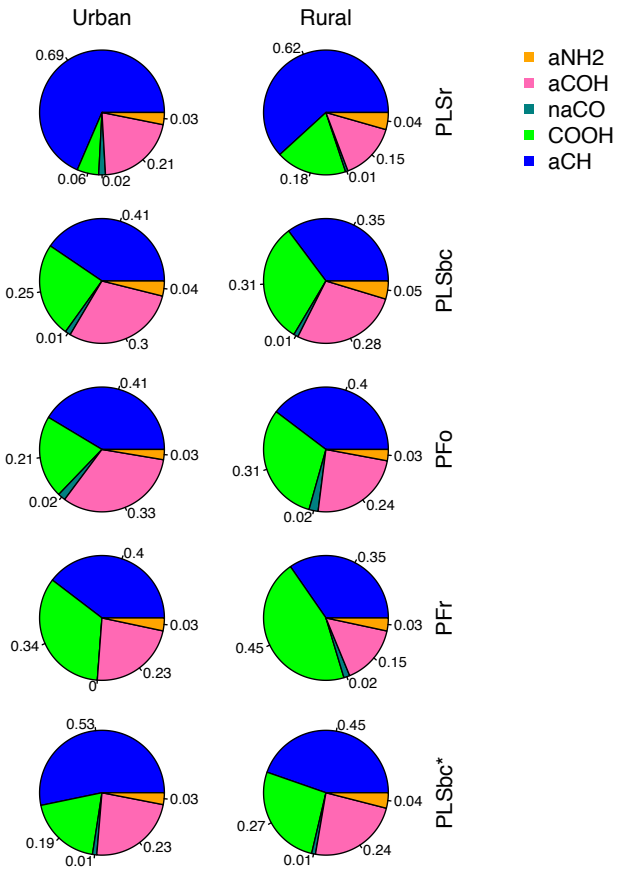

Figure S10: FG distribution in OM. PLSr refers to partial least square using raw spectra. PFr refers to peak-fitting using the recalibrated absorption coefficients. PLSbc refers to partial least square using baseline corrected spectra. PLSbc* referes to the PLSbc with heuristic selection of the number of LVs for aCH.

S8 Variations in FG abundance with number of LVs used in the PLSbc model.

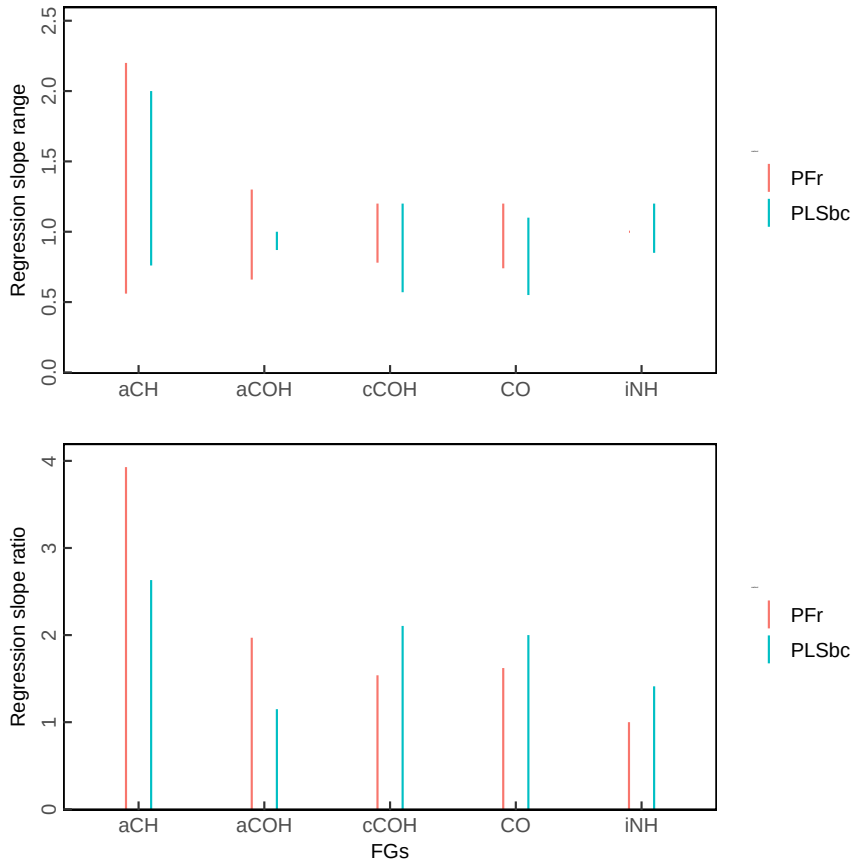

Figure S11: Top row: Range of absorption coefficients relative to selected (for PF) or regression slopes of PLS solutions normalized to selected solution (for PLSbc). Bottom panel: the ratio of maximum to minimum values of top panel for each FG.

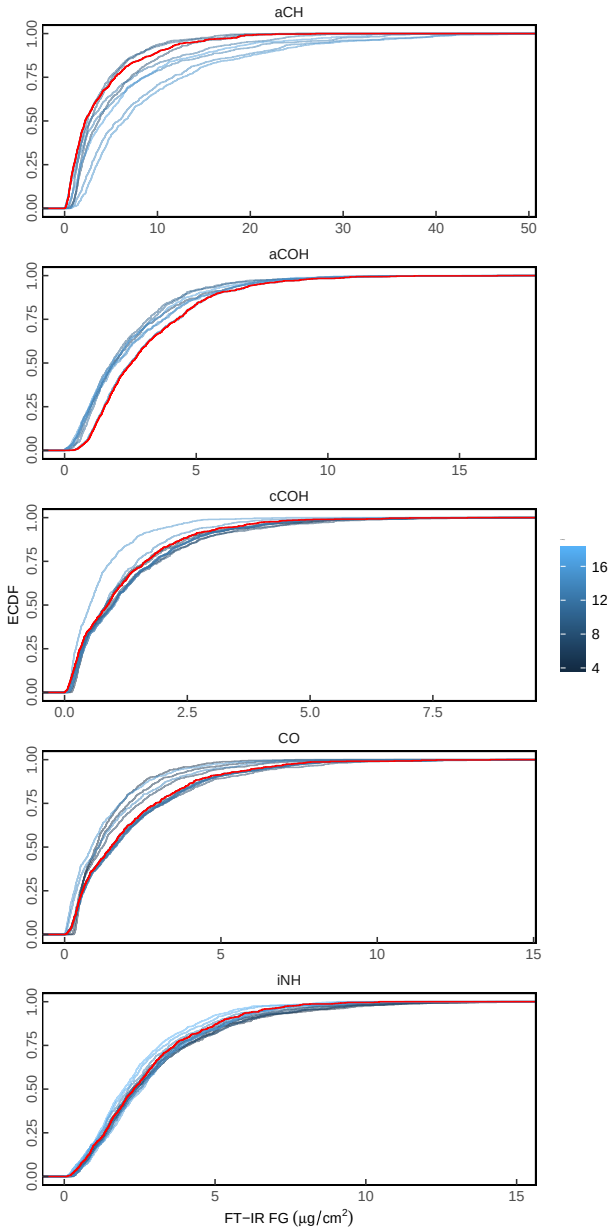

Figure S12: Empirical cumulative density function (ECDF) of the FG abudance predictions for models with different number of LVs which have correlation above 0.95 with the model selected using the method proposed by **?** (red line).

S9 Variable Importance in Projection and Explained Variation in carbonyl calibration models

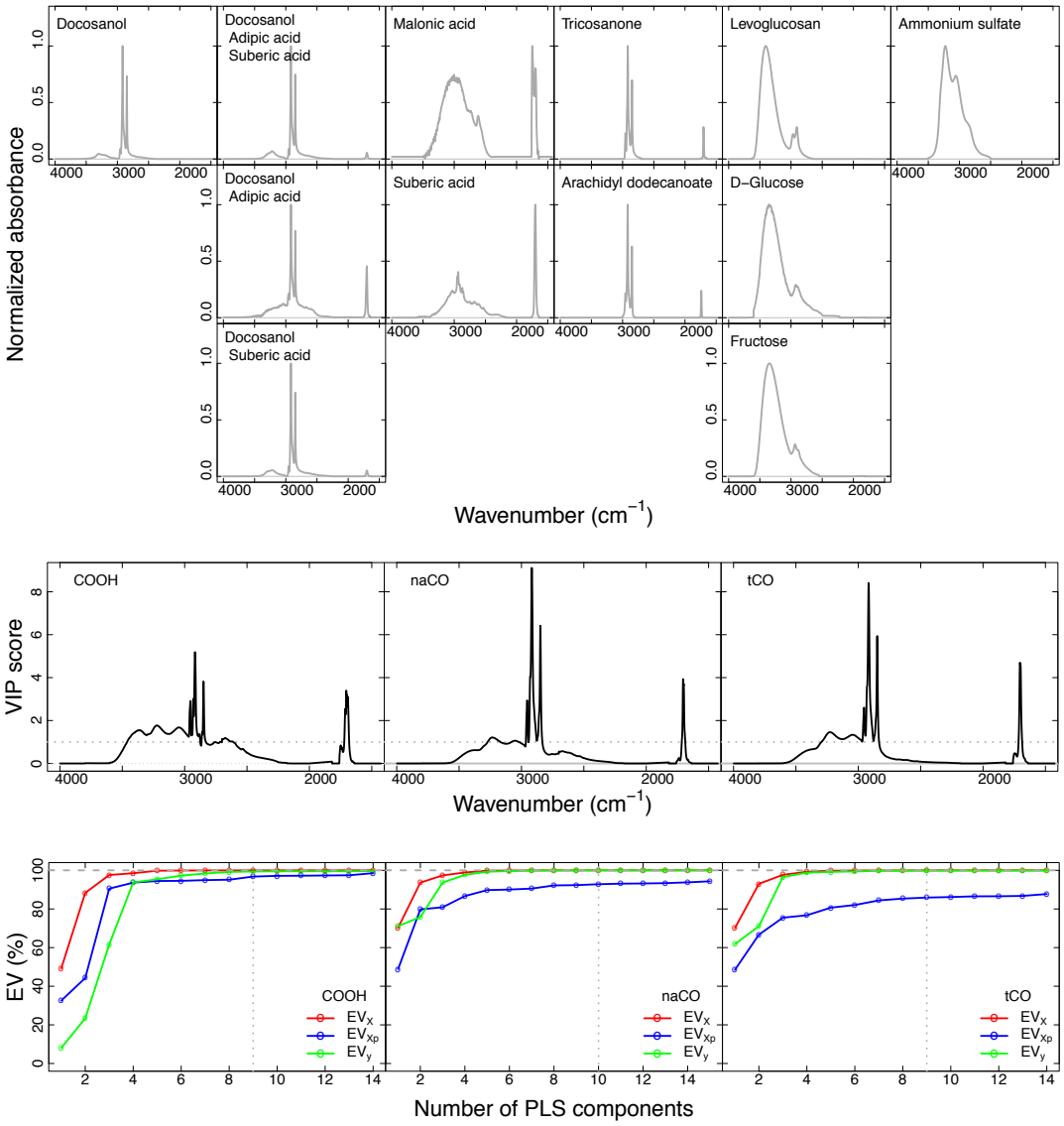

Figure S13: Profiles of laboratory standards, Variable Importance in Projection (VIP) scores for COOH, naCO and tCO, and their cumulative explained variations.

S10 OM/OC probability distributions

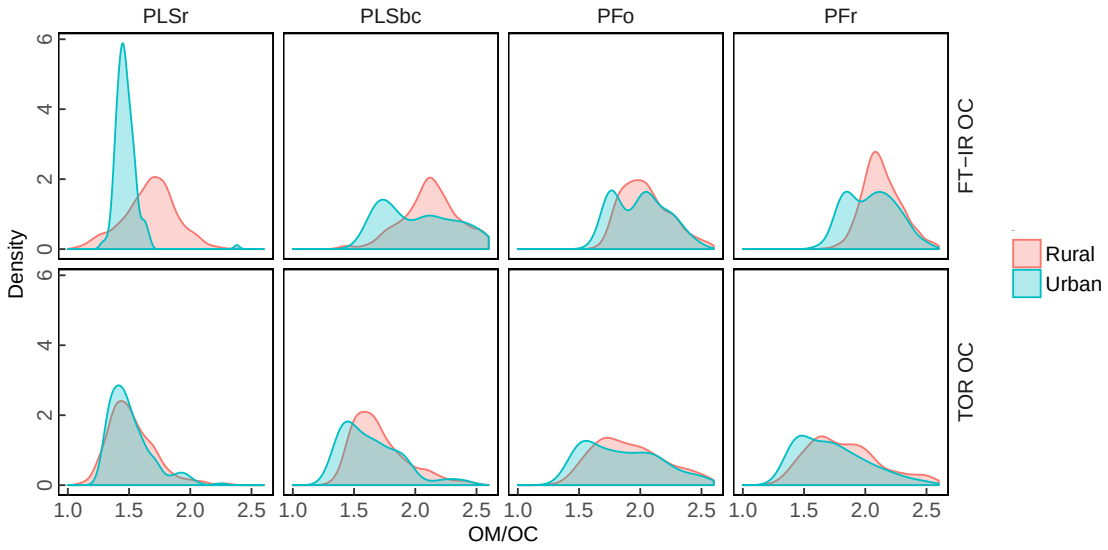

Figure S14: Probability densities shown for OM/OC ratios. Top row is OM/OC estimated solely from FTIR estimates; bottom row is OM/OC using TOR OC for normalization.

S11 Christiansen peak effect in anomalous clusters 19 and 20

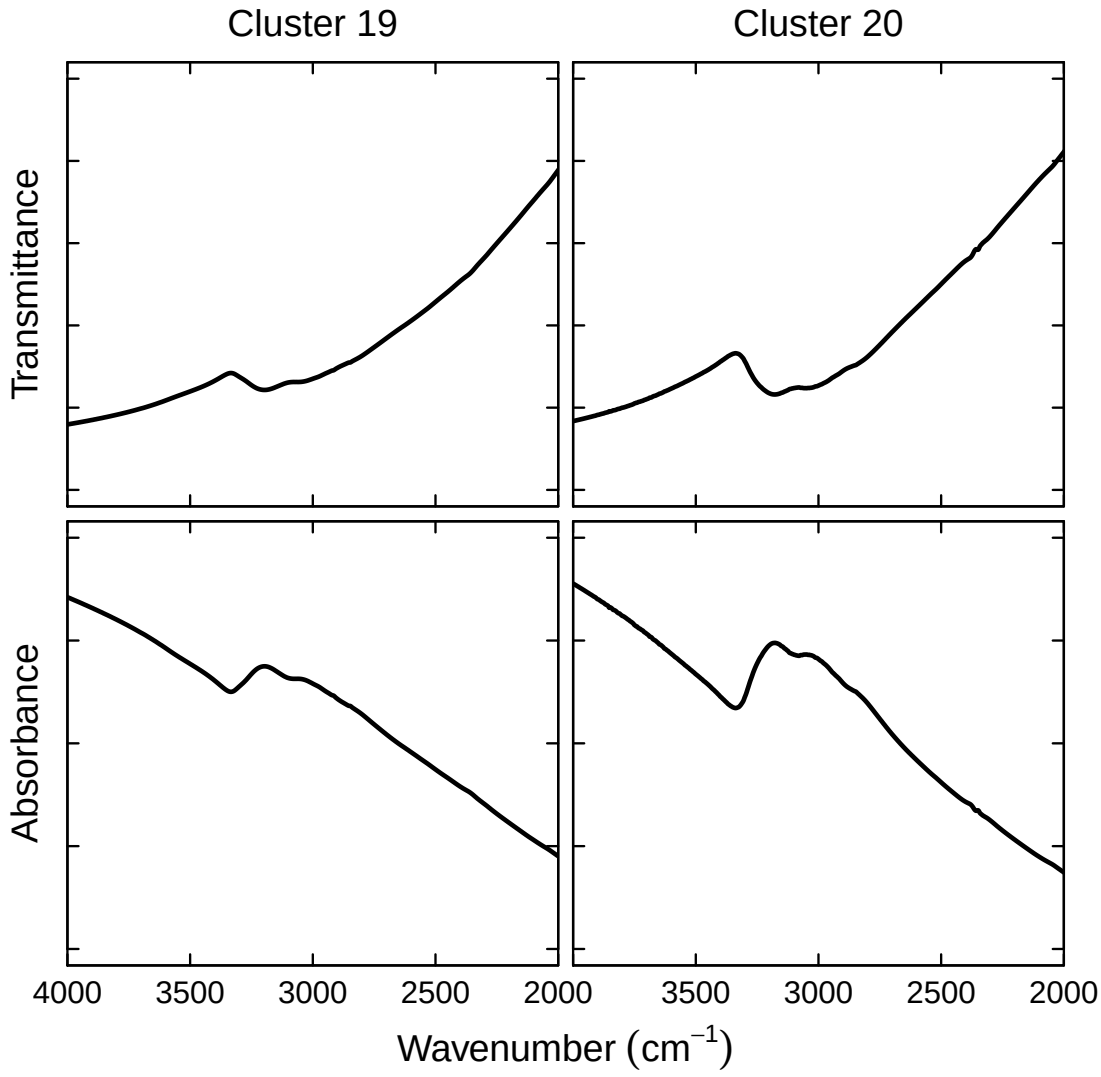

Figure S15: Illustration of samples exhibiting the Christiansen peak effect.

S12 Variations of OC, FG and OM/OC with number of LVs for aCH with the PLSbc model

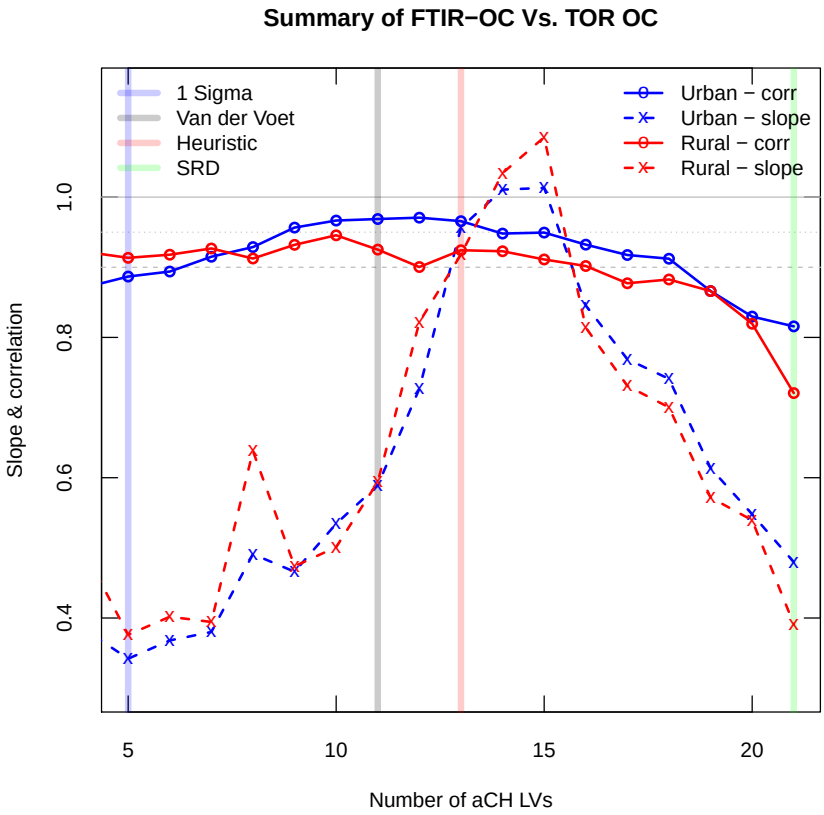

Figure S16: Summary of the comparison of estimated OC (FG OC) gainst OC measured by TOR method (TOR OC) for a different number of aCH Latent variables (LVs) of the PLSbc model. The vertical gray line denotes the number of LVs selected using the method proposed by **?**. The vertical red line denotes the heuristic optimum.