Supplemental Information for "Submicron organic aerosol in Tijuana, Mexico, from local and Southern California sources during the CalMex campaign"

Contents

S1 Dimension reduction methods **1**
- S1.1 PMF description . . . 1
- S1.2 Augmentation of the ACSM PMF standard deviation matrix . . . 2
- S1.3 Comparison of ACSM spectra classification analysis by LDA and k-nn algorithms . . . 2
- S1.4 Using classification analysis to approximate ensemble mixture proportions . . . 3
- S1.5 Additional methods for ACSM spectra analysis: SVD and regression analysis (CMB) . . . 4

S2 Apportionment of OM **4**
- S2.1 Synthesis of FTIR and ACSM PMF solutions . . . 4
- S2.2 Mass fragment and VOC ratios . . . 7
  - S2.2.1 Lag-time correlations . . . 7
  - S2.2.2 Toluene-to-benzene ratios . . . 7

S3 Meteorological analysis **7**

S1 Dimension reduction methods

S1.1 PMF description

Given a column matrix of row vectors (i.e., the sample spectra), $\boldsymbol{X}$, the forward model is expressed as $\boldsymbol{X} = \boldsymbol{GF} + \boldsymbol{E}$. $\boldsymbol{G}$ and $\boldsymbol{F}$ are matrices comprising component strengths and profiles, respectively; $\boldsymbol{E}$ is the residual matrix. The Q (or χ^2) value to be minimized is defined by the canonical objective function $Q = \sum_i \sum_j e_{ij}^2 / s_{ij}^2$, where e_{ij} are the residuals (elements of $\boldsymbol{E}$) and s_{ij} define the weighting for the fit. These weights are derived from mechanistic estimates of the measurement error (Polissar et al., 1998) for both FTIR (Russell et al., 2009b) and ACSM (Ng et al., 2011). The goodness-of-fit metric (Q) is evaluated against a theoretical or expected value, which is approximated by the degrees of freedom in the system: $Q_{expected} \approx \nu = m \cdot n - p(m+n)$ (Paatero et al., 2002). For data sets with large number of n variables (e.g., FTIR spectra) or large number of m samples (e.g., ACSM), is essentially the number of elements in the data matrix, $\boldsymbol{X}$ (Ulbrich et al., 2009). The assumption in this model is that $\boldsymbol{X}$ is composed of a signal and noise, and we wish to represent the signal with factor components and allow the residual term to carry the noise. A commonly prescribed criterion for selecting a solution is that $Q/Q_{expected} \approx 1$, according to the premise that $\boldsymbol{S}$ represents the magnitude of noise component, and factor components are fitting only the true signal such that $\boldsymbol{E} \approx \boldsymbol{S}$.

The four-factor FTIR solution discussed has a $Q/Q_{expected}$ value of 0.8 and is chosen as it satisfies the criteria outlined by Russell et al. (2009b): the factor strengths are not strongly correlated ($|r| < 0.5$) and the reconstructed X matrix adequately reproduces the original spectra (Explained Variation $> 95\%$). Factor components are presumably a mixture of compounds; source classes are inferred for each of these mixtures by examining correlations with elemental tracers and confirmed by comparison to factor spectra obtained from previous campaigns (Russell et al., 2011). For ACSM, we choose a two-factor solution under a modified $\boldsymbol{S}$ matrix (discussed in Section S1.2). The PMF analysis for both FTIR and ACSM is performed using the PMF2 algorithm by Paatero and Tapper (1994). Exploration of parameters and evaluation of solutions are performed using a set of scripts written in R (R Development Core Team, 2012).

S1.2 Augmentation of the ACSM PMF standard deviation matrix

If x = the measured signal (OM), u the true signal, and ϵ represents the instrument noise, a scalar representation of their relationship can be written as

$$x = u + \epsilon \ . \tag{S1}$$

$\mathrm{Var}(\epsilon)$ is estimated from measurement errors compounded for ion counts converted to analog signals (Ng et al., 2011). The PMF statement is equivalent to Equation (S1) except that ϵ represents the residuals (fitting error), which is assumed to be approximately equal to the measurement error in magnitude for the purpose of finding a suitable solution (Paatero and Tapper, 1994). Upon an initial iteration of PMF decomposition, we find that the $Q/Q_{expected}$ does not converge to unity even for a large number of factors (out to 12), suggesting that $\boldsymbol{S}$ may be underestimated for this data set ($Q/Q_{expected}$ is approximately 3 for a four-factor solution). This underestimation leads to solutions in which some of the factors appear to representing additional noise rather than an underlying component of the true signal. To remedy this situation, we select a solution in which the model is purposely over-fitting the data such that the factors with the lowest loadings are fitting the noise. These components are taken as estimates of the unaccounted measurement error and append this to the original $\boldsymbol{S}$ matrix. In formula, this is expressed as $x = u + \delta + \epsilon$. In this solution, a factor component, δ, contributes little to explaining the systematic variation in x (median correlation with any m/z signal $\sim$0.1). Therefore, the measurement error matrix is augmented by assuming that $\mathrm{Var}(\delta) = \delta$ (limit of Poisson statistics) and deriving a new estimate of the standard deviation matrix. In terms of the matrix formulation of PMF, the first decomposition yields $\boldsymbol{X} = (\boldsymbol{GF})_u + (\boldsymbol{GF})_\delta + \boldsymbol{E}$; the standard deviation matrix $\boldsymbol{S}$ is augmented as $\boldsymbol{S}_{(new)} = \sqrt{(\boldsymbol{GF})_\delta^2 + \boldsymbol{S}^2}$. This reduces the Q-value from 3.9 to 1.7 for a two-factor solution, but more importantly, eliminates the generation of solutions in which a δ term or terms are included (i.e., components appear to represent true signals). Residuals (e_{ij}) normalized by the corresponding standard deviation are shown in Figure (S1).

S1.3 Comparison of ACSM spectra classification analysis by LDA and k-nn algorithms

Linear discriminant analysis (LDA) is a method for determining linear decision boundaries (hyperplanes) for multidimensional, continuous variables, to delineate regions of observations belonging to a particular class or category. The k-nearest neighbor (k-nn) method assigns categories to multivariate observations based on proximity (the Euclidean distance metric is most commonly used) to elements in the training set. Details of these statistical learning methods are described by (Hastie et al., 2009).

As described in Section 2, ACSM mass fragment spectra are compared to a training set of unit mass resolution mass spectra from AMS (Ulbrich et al., 2009, 2012). As a projection of the measurements in the space of normalized mass fragment concentration pairs (including f_{43}, f_{44}, f_{57}, f_{60}; each plotted against f_{44} is shown in Figure 7) indicates that a segregation of the training set naturally lends itself to a partitioning of the composition space such a method as linear discriminant analysis (LDA), we can obtain a first-order estimate of the types of samples present. Given the number of assumptions required for LDA (equal covariances, multi-normality) that are difficult to assess for the small number of samples in the training set, k-nearest

neighbor (k-nn) classification is also used to provide another method of classification subject to different a set of constraint criteria. We use a reduced set for the classification based on our understanding that m/z 43, 44, 57, and 60 are the important mass fragments that differentiate among OOA, HOA, and BBOA aerosol (Ng et al., 2011) and avoid using the full spectrum where collinearity may inflate the errors in the classification. We apply a square-root transformation of the feature vector so that normal distributions are approximated by the data, and scale each vector by its 2-norm distance. Both LDA and k-nn classifications are performed in *R* (R Development Core Team, 2012) using the `MASS` and `class` libraries, respectively. Figure S2 shows the similarities in classified fractions according to LDA and k-nn analysis.

S1.4 Using classification analysis to approximate ensemble mixture proportions

Let us assume the existence of two end-member states, M_1 and M_2. x represents the proportion of component M_1 in a sample, obtained from feature vector $\mathbf{v}$ of an arbitrary number of dimensions using a mapping represented by g_{M_1}. If we allow $\mathcal{M}_1$ and $\mathcal{M}_2$ to denote categories with approximately similar composition to their respective end members, we can also define an indicator function $\mathrm{I}_{\mathcal{M}_1}(\mathbf{v})$:

$$\mathrm{I}_{\mathcal{M}_1}(\mathbf{v}) = \begin{cases} 1 & \text{if } \mathbf{v} \in \mathcal{M}_1 \\ 0 & \text{if } \mathbf{v} \notin \mathcal{M}_1 \end{cases} .$$

According to this formulation, the result of classification analysis is to alternatively represent the feature vector for each sample with a discrete (binary) value, x', rather than a continuous variable, x:

$$\begin{aligned} x &= g_{M_1}(\mathbf{v}) \\ x' &= \mathrm{I}_{\mathcal{M}_1}(\mathbf{v}) \ . \end{aligned}$$

The expected value of X' (a Bernoulli random variable) can be shown to approximate the ensemble average of X (a continuous random variable) according to the indicator function. That is to say that we approximate $\mathrm{E}(X)$ with $\mathrm{E}(X')$ by replacing x with x' for each sample i:

$$\frac{1}{n}\sum_{i}^{n} g_{M_1}(\mathbf{v}_i) \approx \frac{1}{n}\sum_{i}^{n} \mathrm{I}_{\mathcal{M}_1}(\mathbf{v}_i) = \frac{n_{\mathcal{M}_1}}{n} \ , \tag{S2}$$

where n is the total number of samples and $n_{\mathcal{M}_1}$ is the number of samples classified into category $\mathcal{M}_1$. The right-hand side of Equation (S2) follows from the fact that the estimator for $\mathrm{E}(X')$ is $\hat{p} = n_{\mathcal{M}_1}/n$. This concept can be extended to extended to N exhaustive end-member states by specification of complementary indicator functions such that

$$\sum_{k=1}^{N} \mathrm{I}_{\mathcal{M}_k}(\mathbf{v}_i) = 1 \ \forall\, i = \{1, \ldots, n\} \ .$$

To fix ideas in a simple, one-dimensional case with two end members, let $x = \mathbf{v}$ and g_{M_1} be the identity function. We define the indicator function in terms of a threshold value, $\phi_{\mathcal{M}_1}$, such that $\mathrm{I}_{\mathcal{M}_1}(x) = \Phi(-x + \phi_{\mathcal{M}_1})$. Φ is the Heaviside unit step function; $\phi_{\mathcal{M}_1}$ is the threshold below which x is categorized as $\mathcal{M}_1$ (and above as $\mathcal{M}_2$). The probabiliy distribution of the Bernoulli random variable, X', is related to the continuous distribution of the original variable such that

$$\mathrm{E}(X') = \Pr(X' = 1) = \Pr(X \in \mathcal{M}_1) = \Pr(X \leq \phi_{\mathcal{M}_1})$$

since the empirical cumulative distribution function (c.d.f.) of X also has the form $\Pr(X \leq \phi) = \Phi(-x + \phi)$ where ϕ is any value in the domain of X, which is [0,1]. Markov's inequality (Wasserman, 2010) can approximately constrain the complementary c.d.f. for a continuous distribution from its expected value. Since in our example the c.d.f. is equal to $\mathrm{E}(X')$, this inequality can give us some indication of how $\mathrm{E}(X)$ and $\mathrm{E}(X')$ are related in this specific case. If we let $\Delta = \left|\mathrm{E}(X') - \mathrm{E}(X)\right|$, then $\Delta \leq \mathrm{E}(X')\left[1 + \phi_{\mathcal{M}_1}\right] - \phi_{\mathcal{M}_1}$,

and we can bound the magnitude of the approximation (between $\mathrm{E}(X)$ and $\mathrm{E}(X')$) for choice of $\phi_{\mathcal{M}_1}$ and estimated values of $\mathrm{E}(X')$ (and also minimize Δ with respect to $\phi_{\mathcal{M}_1}$). While many classification algorithms preclude definition of a simple threshold and each estimation method may require different prepration of feature vectors to satisfy required assumptions, this illustration shows a very simple case in which discrete categorizations can be used to approximate mixture proportions. In the more general case, if we define $d = I_{\mathcal{M}_1}(\mathbf{v}) - g_{M_1}(\mathbf{v})$, then $\Delta = |\mathrm{E}(D)|$ (from Equation S2). Therefore, we expect that the bias in estimation between the two methods would depend on the distribution of D.

It should be noted that the correctness of the approximation strongly depends on the classification scheme which defines the indicator functions. Therefore, the solution of 88/12% OOA/HOA (Section S1.5) rather than 60/40% OOA-1/OOA-2 (Section 3.2) apportionment cannot be ruled out, especially if the feature vectors or classification algorithm cannot fully discriminate between HOA and OOA-2 if they are approximately collinear along a continuum of oxidation states. But as stated in Section 3.2, this partitioning affects the estimation of the more and less oxygenated fractions, but their respective trends are robust with respect to either interpretation. Therefore, the source association of the components based on the diurnal variations showed in Figure 8 remains unchanged.

S1.5 Additional methods for ACSM spectra analysis: SVD and regression analysis (CMB)

Singular Value Decomposition (SVD) on the sample matrix $\boldsymbol{X}$ (which does not account for the measurement error matrix, $\boldsymbol{S}$) indicates that two components explain approximately 80% of the variation in the measurements, consistent with an Explained Variation of 80% from PMF analysis for a two-component solution.

Regression analysis, often referred to as a Chemical Mass Balance (CMB) approach (Chow and Watson, 2002; Ng et al., 2011), to component apportionment is also used but yielded results not consistent with the domain delineated by Figure 7 because of the large collinearity among regressands. Approximate apportionment to HOA and OOA (HOA_{est} and OOA_{est}, respectively) from unit-mass resolution AMS measurements using OM-equivalent concentrations of mass fragments 44 and 57 (C_{44} and C_{57}, respectively) is suggested by Ng et al. (2011):

$$\begin{pmatrix} HOA_{\mathrm{est}} \\ OOA_{\mathrm{est}} \end{pmatrix} = \begin{pmatrix} 0 & b & a \cdot b \\ d & 0 & c \end{pmatrix} \begin{pmatrix} 1 \\ C_{57} \\ C_{44} \end{pmatrix}. \tag{S3}$$

Using Equation (S3) with coefficients $\{a = 0.095, b = 15.2, c = 6.92, d = 0.07\}$ estimated from median values of various campaigns (Ng et al., 2011) provides an alternate method of partitioning the OM into more and less oxygenated fractions. The HOA and OOA apportioned using this method correlate well with OOA-2 (r=0.86) and OOA-1 (r=0.77), respectively. Therefore, conclusions dependent on the relative variation of the less oxygenated and more oxygenated components in time are robust with respect to either interpretation. However, the regression coefficients β={intercept,slope} are {-0.25,0.46} for HOA_{est} on OOA-2, and {-0.5,1.8} for OOA_{est} on OOA-1. This alternative interpretation would suggest that the transported fraction to OM may be as high as 88% – rather than 60% as suggested using the OOA-1 and OOA-2 PMF solution – based on the justified assumption that the majority of the more oxygenated fraction of OM is due to long-range transport (Section 3.3). However, more weight is placed on the OOA-1 and OOA-2 PMF solution as it is derived for measurements at the Tijuana site specifically, and the solution is supported by the chemical composition space delineated by Figure 7 and classification analysis discussed in section S1.3.

S2 Apportionment of OM

S2.1 Synthesis of FTIR and ACSM PMF solutions

As the FTIR and ACSM provides complementary information regarding the organic fraction of ambient aerosols, we wish to harmonize our interpretation of our measurements by combining the information re-

trieved from both instruments. One method is to combine measurements from both instruments into a single data matrix for multivariate PMF analysis, which poses many challenges. For instance, the number of variables provided by each instrument are different; appropriate downweighting must be applied to reduce redundancies for each instrument, and for one set of measurement with respect to the other such that one is not favored more heavily in the solution. These issues are effectively handled through the adjustment of the standard deviation matrix; a bottom-up approach to specification of this matrix, including the measurement-uncertainty component, requires absolute calibration such that they are comparable across instruments. Slowik et al. (2010) combined measurements from an AMS and PTR-MS to find covarying mass-fragments from PMF analysis, and iteratively adjusted the elements of the standard deviation matrix (collectively by instrument) with a single scaling factor to obtain empirical weightings such that mass fragments from both instruments were well represented and reproduced by the solutions. In this way, Slowik et al. (2010) retained relative weights (due to measurement uncertainties) for each instrument, but adjusted overall according to their information content with respect to the other instrument. Combining absorbance and mass fragment spectra for PMF analysis could possibly be approched in this way to ensure that solutions represent apportionment based on variation, rather than (possibly miscalibrated) variance. However, in this work, we adopt a naive approach which provides an interpretation of the two measurements after PMF decomposition has been applied independently.

Taking a rather extreme perspective, we assume FTIR PMF components retain information regarding source class (Russell et al., 2011) and ACSM PMF components indicate information regarding aerosol age (Jimenez et al., 2009). From this analysis we attempt to approximate age of OM attributed to each source class (from which notions of local and remote origins emerge). Let OM_A and OM_F represent OM concentrations measured by ACSM and FTIR, respectively, and OM_C represent the OM concentration measured by both techniques (for a single sample). Disregarding non-systematic disturbances to each set of measurements (PMF presumably retains the signal to factor components and apportions these disturbances to the residual term), it is assumed that the primary differences among these metrics of OM are the unmeasured compounds or organic aerosol fractions due to reasons described in Section 3.2. $\boldsymbol{\alpha}$ is a vector which indicates the fraction of OM attributed to components $\mathcal{A} = \{\text{OOA-1}, \text{OOA-2}\}$ from ACSM PMF analysis, or components $\mathcal{F} = \{\text{FF1}, \text{FF2}, \text{BB}\}$ from FTIR PMF analysis (as denoted by subscripts). The Marine factor is excluded from this portion of the analysis as we expect marine organic OM associated with non-refractory sea salt not to be measured by the ACSM (Section 3.2). We additionally define a vector of all ones, $\mathbf{1}_m$, where m indicates the number of elements (often defined here by the cardinality of a set, $|\cdot|$, or number of samples, n). Using these definitions, OM_C can be defined in terms of the fractional sum of components for each instrument: $OM_C = OM_A \boldsymbol{\alpha}_{\mathcal{A}}^T \mathbf{1}_{|\mathcal{A}|} = OM_F \boldsymbol{\alpha}_{\mathcal{F}}^T \mathbf{1}_{|\mathcal{F}|}$. In this context, $\mathbf{1}_m$ is introduced as a postfix operator which sums elements of a vector (or rows or columns of matrices, depending on orientation). In our present specification, $\boldsymbol{\alpha}_{\mathcal{A}}^T \mathbf{1}_{|\mathcal{A}|}$ and $\boldsymbol{\alpha}_{\mathcal{F}}^T \mathbf{1}_{|\mathcal{F}|}$ are not required to equal unity, as the marine component for FTIR has been excluded from $\boldsymbol{\alpha}_{\mathcal{F}}$, for instance. We can eliminate the explicit references to OM by redefining $\boldsymbol{\alpha}_{\mathcal{A}}$ and $\boldsymbol{\alpha}_{\mathcal{F}}$ with scaling factors OM_A/OM_C and OM_F/OM_C, respectively, such that

$$1 = \boldsymbol{\alpha}_{\mathcal{A}}^T \mathbf{1}_{|\mathcal{A}|} = \boldsymbol{\alpha}_{\mathcal{F}}^T \mathbf{1}_{|\mathcal{F}|} \tag{S4}$$

The proportion of OM_C attributed to any component from one measurement can be described with respect to contributions from the other, which follows from Equation (S4):

$$\begin{aligned} \boldsymbol{\alpha}_{\mathcal{A}} &= \boldsymbol{\alpha}_{\mathcal{A}} \boldsymbol{\alpha}_{\mathcal{F}}^T \mathbf{1}_{|\mathcal{F}|} \\ \boldsymbol{\alpha}_{\mathcal{F}} &= \boldsymbol{\alpha}_{\mathcal{F}} \boldsymbol{\alpha}_{\mathcal{A}}^T \mathbf{1}_{|\mathcal{A}|} \ . \end{aligned} \tag{S5}$$

The two outer products, $\boldsymbol{\alpha}_{\mathcal{A}} \boldsymbol{\alpha}_{\mathcal{F}}^T$ and $\boldsymbol{\alpha}_{\mathcal{F}} \boldsymbol{\alpha}_{\mathcal{A}}^T$, define the fraction of FTIR components associated with ACSM components, and vice versa. For instance, the former case (Equation S5) is also written,

$$\begin{pmatrix} \alpha_{\text{OOA-1}} \\ \alpha_{\text{OOA-2}} \end{pmatrix} = \begin{pmatrix} \alpha_{\text{OOA-1}}\alpha_{\text{FF1}} & \alpha_{\text{OOA-1}}\alpha_{\text{FF2}} & \alpha_{\text{OOA-1}}\alpha_{\text{BB}} \\ \alpha_{\text{OOA-2}}\alpha_{\text{FF1}} & \alpha_{\text{OOA-2}}\alpha_{\text{FF2}} & \alpha_{\text{OOA-2}}\alpha_{\text{BB}} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \ , \tag{S6}$$

where the first matrix on the right-hand side is $\boldsymbol{\alpha}_{\mathcal{A}}\boldsymbol{\alpha}_{\mathcal{F}}^T$. Campaign-averaged contributions for $OM_C\boldsymbol{\alpha}_{\mathcal{A}}\boldsymbol{\alpha}_{\mathcal{F}}^T$ are shown in Figure S3. Averaged quantities of $\boldsymbol{\alpha}_{\mathcal{A}}\boldsymbol{\alpha}_{\mathcal{F}}^T$ and $\boldsymbol{\alpha}_{\mathcal{F}}\boldsymbol{\alpha}_{\mathcal{A}}^T$ are essentially a measure of the uncentered covariance scaled by individual measurements of OM. To illustrate this interpretation, let us define two matrices, $\boldsymbol{A}_{\mathcal{A}}$ and $\boldsymbol{A}_{\mathcal{F}}$ for ACSM and FTIR measurements, respectively, which contain all rescaled values of $\boldsymbol{\alpha}$ for the campaign:

$$\boldsymbol{A}_{\mathcal{A}} = \begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{A},1} & \boldsymbol{\alpha}_{\mathcal{A},2} & \dots & \boldsymbol{\alpha}_{\mathcal{A},n}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\alpha}_{\text{OOA-1}}^T \\ \boldsymbol{\alpha}_{\text{OOA-2}}^T\end{pmatrix} \quad \text{and} \quad \boldsymbol{A}_{\mathcal{F}} = \begin{pmatrix}\boldsymbol{\alpha}_{\mathcal{F},1} & \boldsymbol{\alpha}_{\mathcal{F},2} & \dots & \boldsymbol{\alpha}_{\mathcal{F},n}\end{pmatrix} = \begin{pmatrix}\boldsymbol{\alpha}_{\text{FF1}}^T \\ \boldsymbol{\alpha}_{\text{FF2}}^T \\ \boldsymbol{\alpha}_{\text{BB}}^T\end{pmatrix} .$$

Each matrix of $\boldsymbol{A}$ is defined in two equivalent representations, either as 1) vectors containing component fractions, $\boldsymbol{\alpha}_{\mathcal{A},i}$ or $\boldsymbol{\alpha}_{\mathcal{F},i}$ – with second indices indicating the sample number – or 2) vectors containing OM mass fractions for each PMF component in $\mathcal{A}$ or $\mathcal{F}$, with number of dimensions equal to the number of measurements, n, in the campaign. We adopt the second notation for illustrating similarities to the array notation of $\boldsymbol{\alpha}_{\mathcal{A}}\boldsymbol{\alpha}_{\mathcal{F}}^T$ displayed in Equation (S6). We can restate the equality in OM reconstruction for each sample (Equation S4):

$$\mathbf{1}_n = \boldsymbol{A}_{\mathcal{A}}^T\mathbf{1}_{|\mathcal{A}|} = \boldsymbol{A}_{\mathcal{F}}^T\mathbf{1}_{|\mathcal{F}|} ,$$

and define their averages:

$$\frac{1}{n}\boldsymbol{A}_{\mathcal{A}}\mathbf{1}_n = \frac{1}{n}\boldsymbol{A}_{\mathcal{A}}\boldsymbol{A}_{\mathcal{F}}^T\mathbf{1}_{|\mathcal{F}|} \tag{S7}$$

$$\frac{1}{n}\boldsymbol{A}_{\mathcal{F}}\mathbf{1}_n = \frac{1}{n}\boldsymbol{A}_{\mathcal{F}}\boldsymbol{A}_{\mathcal{A}}^T\mathbf{1}_{|\mathcal{A}|} . \tag{S8}$$

A product of $\mathbf{A}_{\mathcal{A}}$ and $\mathbf{A}_{\mathcal{F}}$ will yield uncentered covariances across components (scaled by the number of observations, n):

$$\frac{1}{n}\boldsymbol{A}_{\mathcal{A}}\boldsymbol{A}_{\mathcal{F}}^T = \frac{1}{n}\begin{pmatrix}\boldsymbol{\alpha}_{\text{OOA-1}}^T\boldsymbol{\alpha}_{\text{FF1}} & \boldsymbol{\alpha}_{\text{OOA-1}}^T\boldsymbol{\alpha}_{\text{FF2}} & \boldsymbol{\alpha}_{\text{OOA-1}}^T\boldsymbol{\alpha}_{\text{BB}} \\ \boldsymbol{\alpha}_{\text{OOA-2}}^T\boldsymbol{\alpha}_{\text{FF1}} & \boldsymbol{\alpha}_{\text{OOA-2}}^T\boldsymbol{\alpha}_{\text{FF2}} & \boldsymbol{\alpha}_{\text{OOA-2}}^T\boldsymbol{\alpha}_{\text{BB}}\end{pmatrix} .$$

Russell et al. (2009a) explored linear relationships between fragments and bonds measured by AMS and FTIR, respectively, through ordinary regression. We consider this approach for explaining variations in $\boldsymbol{\alpha}$s in $\mathcal{A}$ through linear combinations of $\boldsymbol{\alpha}$s in $\mathcal{F}$. We extend individual statements of regression to a multivariate, multiple regression expression, $\boldsymbol{A}_{\mathcal{A}} = \boldsymbol{\beta}^T\boldsymbol{A}_{\mathcal{F}} + \boldsymbol{E}$, where $\boldsymbol{E}$ is the residual matrix. The expected value of $\boldsymbol{A}_{\mathcal{A}}$ can be written as $\hat{\boldsymbol{A}}_{\mathcal{A}} = \boldsymbol{A}_{\mathcal{A}}\boldsymbol{H}^T$. A possible solution for the regression coefficients are $\hat{\boldsymbol{\beta}} = \left(\boldsymbol{A}_{\mathcal{F}}\boldsymbol{A}_{\mathcal{F}}^T\right)^{-1}\boldsymbol{A}_{\mathcal{F}}\boldsymbol{A}_{\mathcal{A}}^T$, and the hat (projection) matrix can be defined as $\boldsymbol{H} = \boldsymbol{A}_{\mathcal{F}}^T\hat{\boldsymbol{\beta}}\left(\boldsymbol{A}_{\mathcal{A}}^T\right)^{-1} = \boldsymbol{A}_{\mathcal{F}}^T\left(\boldsymbol{A}_{\mathcal{F}}\boldsymbol{A}_{\mathcal{F}}^T\right)^{-1}\boldsymbol{A}_{\mathcal{F}}$. Of course, a naive solution is proposed for illustration in this case, but non-negativity should be considered in the actual specification of $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{H}$. To summarize the two approaches based on Equation S8 and Russell et al. (2009a), we can postfix $\boldsymbol{A}_{\mathcal{A}} = \left(\begin{smallmatrix}\boldsymbol{\alpha}_{\text{OOA-1}}^T \\ \boldsymbol{\alpha}_{\text{OOA-2}}^T\end{smallmatrix}\right)$ with either $\frac{1}{n}\boldsymbol{A}_{\mathcal{F}}\mathbf{1}_{|\mathcal{F}|}$ to get $\left(\begin{smallmatrix}\hat{\alpha}_{\text{OOA-1}} \\ \hat{\alpha}_{\text{OOA-2}}\end{smallmatrix}\right)$, or with $\boldsymbol{H}^T$ to get $\left(\begin{smallmatrix}\hat{\boldsymbol{\alpha}}_{\text{OOA-1}}^T \\ \hat{\boldsymbol{\alpha}}_{\text{OOA-2}}^T\end{smallmatrix}\right)$. The former statement makes a remark regarding the marginal expectation $\mathrm{E}(X)$ from the joint expectation $\mathrm{E}(X, Y)$, and the latter, the conditional expectation $\mathrm{E}(X|Y = y)$, where X and Y are used to denote any pair of covariates in standard scalar notation. The relationship between $\boldsymbol{A}_{\mathcal{A}}$ and $\boldsymbol{A}_{\mathcal{F}}$ are embodied in $\boldsymbol{A}_{\mathcal{A}}\boldsymbol{A}_{\mathcal{F}}^T$ in the former case, and $\hat{\boldsymbol{\beta}}$ (matrix of regression coefficients) for the latter case.

Since the separation between FTIR FF1 and FF2 are thought to be in degree of oxygenation (Russell et al., 2011), the expectation is that FF1 and FF2 are less and more aged, respectively, which is not supported by this analysis if OOA-1 and OOA-2 are strictly interpreted as indicators of age in Tijuana. It is therefore possible that there are other factors which lead to the separation of ACSM OM into OOA-1 and OOA-2, and FTIR OM into FF1 and FF2 at this location. For instance, there may be aspects of molecular composition that are similar (from the perspective of the ACSM mass spectra) across different source types, and therefore an overlap in functional groups apportioned to OOA-1 and OOA-2. Using the same method

(Equation **??**) but letting $\boldsymbol{\alpha}_{\mathcal{F}}$ represent FTIR OFG fractions rather than PMF factor fractions, we can estimate OFG contributions (sans marine OM) to the different ACSM PMF factors and provide support for this interpretation (Figure S4). As discussed in Section 3.2, another interpretation (not mutually exclusive with respect to the previous statement) is that some fraction of FF2 is associated with non-refractory mineral dust (Section 3.2). OM associated with this material is not believed to be sampled by ACSM. Regarding the separation of other components (e.g., BB) between OOA-1 and OOA-2 fractions, the coarse time resolution of the FTIR measurements may also affect the ability to resolve component contributions well (Henry, 2003).

Liggio et al. (2010) apportioned HOA to primary (local) OM, and OOA to a combination of aged backgroumd OM and locally-produced SOA. In our case, we conclude that local production of SOA was small at our measurement site, based on estimates of diurnal increases in OM during photochemically active periods (approximately noon each day) and consistently high degree of oxygenation (Sections 3.2 and 3.3). With similar reasoning, we consider our less oxygenated aerosol (OOA-2) to local aerosol and more oxygenated (OOA-1) as aged. As a zeroth-order estimate, we disregard the separation in fossil fuel combustion OM by FTIR PMF into FF1 and FF2 and consider this as a single source component, and evaluate the local and regional contributions based on the degree of oxygenation of mass fragment spectra (Jimenez et al., 2009). In this case, 60% of the FF OM is associated with OOA-1 (50% if we disregard FF2 contribution to OM measured by FTIR in accordance with our previously stated hypothesis); this value may be considered a lower bound on the average contribution of anthropogenic combustion-related material transported to Tijuana (and oxygenated in the process) from regional sources (Section 3.3).

S2.2 Mass fragment and VOC ratios

S2.2.1 Lag-time correlations

The hourly lag time correlations between ACSM mass fragments and selected VOCs from PTR-MS measurements, as discussed in Section 3.3, are shown in Figure S5.

S2.2.2 Toluene-to-benzene ratios

The toluene-to-benzene ratio, $R_{[T/B]}$, is used as a metric of age and a parameter by which oxygenated aerosol is fractioned into primary and secondary, and background OM (e.g., Liggio et al., 2010). While requirements for chemical specificity precludes apportionment using ACSM mass spectra, correlations of $R_{[T/B]}$ with O/C ratio, and OOA-1 and OOA-2 normalized by OM (Figure S6) suggests that constant and point-source emission assumption (Liggio et al., 2010) required for $R_{[T/B]}$ analysis are not likely to be valid. Despite the constant wind direction/ direction of origin of airmasses to Tijuana, it may be the case that that mixing of airmasses or the existence of multiple sources along the trajectory path leads to inconsistent proportions of toluene to benzene with which to measure airmass age.

S3 Meteorological analysis

Wind speed are estimated to vary between 2-5 m/s as calculated by the HYSPLIT model (Figure S7) streamlines with end-heights of 10, 50, and 100 m as described in Section 2. Using these values, The aged aerosol formed $\geq$10 hours prior is determined to lie beyond the Tijuana and San Diego regions (Figure S8).

References

Chow, J. C., Watson, J. G., Mar. 2002. Review of pm2.5 and pm10 apportionment for fossil fuel combustion and other sources by the chemical mass balance receptor model. Energy & Fuels 16 (2), 222–260.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The elements of statistical learning: data mining, inference, and prediction. Springer Verlag.

Henry, R. C., 2003. Multivariate receptor modeling by n-dimensional edge detection. Chemometrics and Intelligent Laboratory Systems 65 (2), 179–189.

Jimenez, J. L., Canagaratna, M. R., Donahue, N. M., Prevot, A. S. H., Zhang, Q., Kroll, J. H., DeCarlo, P. F., Allan, J. D., Coe, H., Ng, N. L., Aiken, A. C., Docherty, K. S., Ulbrich, I. M., Grieshop, A. P., Robinson, A. L., Duplissy, J., Smith, J. D., Wilson, K. R., Lanz, V. A., Hueglin, C., Sun, Y. L., Tian, J., Laaksonen, A., Raatikainen, T., Rautiainen, J., Vaattovaara, P., Ehn, M., Kulmala, M., Tomlinson, J. M., Collins, D. R., Cubison, M. J., Dunlea, E. J., Huffman, J. A., Onasch, T. B., Alfarra, M. R., Williams, P. I., Bower, K., Kondo, Y., Schneider, J., Drewnick, F., Borrmann, S., Weimer, S., Demerjian, K., Salcedo, D., Cottrell, L., Griffin, R., Takami, A., Miyoshi, T., Hatakeyama, S., Shimono, A., Sun, J. Y., Zhang, Y. M., Dzepina, K., Kimmel, J. R., Sueper, D., Jayne, J. T., Herndon, S. C., Trimborn, A. M., Williams, L. R., Wood, E. C., Middlebrook, A. M., Kolb, C. E., Baltensperger, U., Worsnop, D. R., Dec. 2009. Evolution of organic aerosols in the atmosphere. Science 326 (5959), 1525–1529.

Liggio, J., Li, S.-M., Vlasenko, A., Sjostedt, S., Chang, R., Shantz, N., Abbatt, J., Slowik, J. G., Bottenheim, J. W., Brickell, P. C., Stroud, C., Leaitch, W. R., Nov. 2010. Primary and secondary organic aerosols in urban air masses intercepted at a rural site. Journal of Geophysical Research-atmospheres 115, D21305.

Ng, N. L., Canagaratna, M. R., Jimenez, J. L., Zhang, Q., Ulbrich, I. M., Worsnop, D. R., Feb. 2011. Real-time methods for estimating organic component mass concentrations from aerosol mass spectrometer data. Environmental Science & Technology 45 (3), 910–916.

Paatero, P., Hopke, P. K., Song, X. H., Ramadan, Z., 2002. Understanding and controlling rotations in factor analytic models. Chemometrics and Intelligent Laboratory Systems 60 (1-2), 253–264.

Paatero, P., Tapper, U., 1994. Positive matrix factorization - a nonnegative factor model with optimal utilization of error-estimates of data values. Environmetrics 5 (2), 111–126.

Polissar, A. V., Hopke, P. K., Paatero, P., 1998. Atmospheric aerosol over alaska - 2. elemental composition and sources. Journal of Geophysical Research-atmospheres 103 (D15), 19045–19057.

R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.
URL http://www.R-project.org/

Russell, L. M., Bahadur, R., Hawkins, L. N., Allan, J., Baumgardner, D., Quinn, P. K., Bates, T. S., 2009a. Organic aerosol characterization by complementary measurements of chemical bonds and molecular fragments. Atmospheric Environment 43 (38), 6100–6105.

Russell, L. M., Bahadur, R., Ziemann, P. J., 2011. Identifying organic aerosol sources by comparing functional group composition in chamber and atmospheric particles. Proceedings of the National Academy of Sciences of the United States of America 108 (9), 3516–3521.

Russell, L. M., Takahama, S., Liu, S., Hawkins, L. N., Covert, D. S., Quinn, P. K., Bates, T. S., 2009b. Oxygenated fraction and mass of organic aerosol from direct emission and atmospheric processing measured on the r/v ronald brown during texaqs/gomaccs 2006. Journal of Geophysical Research-atmospheres 114.

Slowik, J. G., Stroud, C., Bottenheim, J. W., Brickell, P. C., Chang, R. Y. . W., Liggio, J., Makar, P. A., Martin, R. V., Moran, M. D., Shantz, N. C., Sjostedt, S. J., van Donkelaar, A., Vlasenko, A., Wiebe, H. A., Xia, A. G., Zhang, J., Leaitch, W. R., Abbatt, J. P. D., 2010. Characterization of a large biogenic secondary organic aerosol event from eastern canadian forests. Atmospheric Chemistry and Physics 10 (6), 2825–2845.

Ulbrich, I., Lechner, M., Jimenez, J., 2012. AMS Spectral Database.
URL http://cires.colorado.edu/jimenez-group/AMSsd/

Ulbrich, I. M., Canagaratna, M. R., Zhang, Q., Worsnop, D. R., Jimenez, J. L., 2009. Interpretation of organic components from positive matrix factorization of aerosol mass spectrometric data. Atmospheric Chemistry and Physics 9 (9), 2891–2918.

Wasserman, L., 2010. All of Statistics: A Concise Course in Statistical Inference (Springer Texts in Statistics). Springer.

Supplemental Tables

Table S1: Summary of instruments used in this work. Further details are described in Section 2.

Analytical method	Measurement	Time resolution
FTIR	PM_1 total OM and organic functional group abundance	6-12 hours
ACSM	Nonrefractory (NR) PM_1 mass (SO_4, NO_3, NH_4, Cl, Organics) and organic molecular mass fragments	15-30 minutes
XRF	PM_1 elemental composition	6-12 hours
PTR-MS	VOC compound concentrations from mass fragment analysis	10 minutes
SP2	Size-resolved black carbon number and mass concentrations (only total number concentrations used in this work)	Single-particle
STXM-NEXAFS	Individual Morphology and composition of carbonaceous particles	Single-particle

Table S2: Organic aerosol spectra components or classes derived for FTIR absorbance spectra and ACSM mass fragment spectra. Further details are described in Section 3 and text of this document.

Instrument	Method	Components or classes (campaign average ± standard deviation, or range in estimated value from multiple methods)
FTIR	PMF	FF1 (40±28%), FF2 (17±19%), BB (20±20%), Marine (23±24%)
FTIR	Nonlinear regression	hydroxyl (22±13%), alkane (44±9%), carboxylic acid (26±7%), ketonic carbonyl (3±5%), primary amine (5±2%), and organic nitrate (0.3±0.3%) functional groups
ACSM	PMF	OOA-1 (60±19%), OOA-2 (40±19%)
ACSM	Linear regression	OOA (88±19%), HOA(12±19%)
ACSM	Classification	OOA-1 (69-77%), OOA-2 (20-26%), BBOA (2-4%), HOA (1-2%)

Supplemental Figures

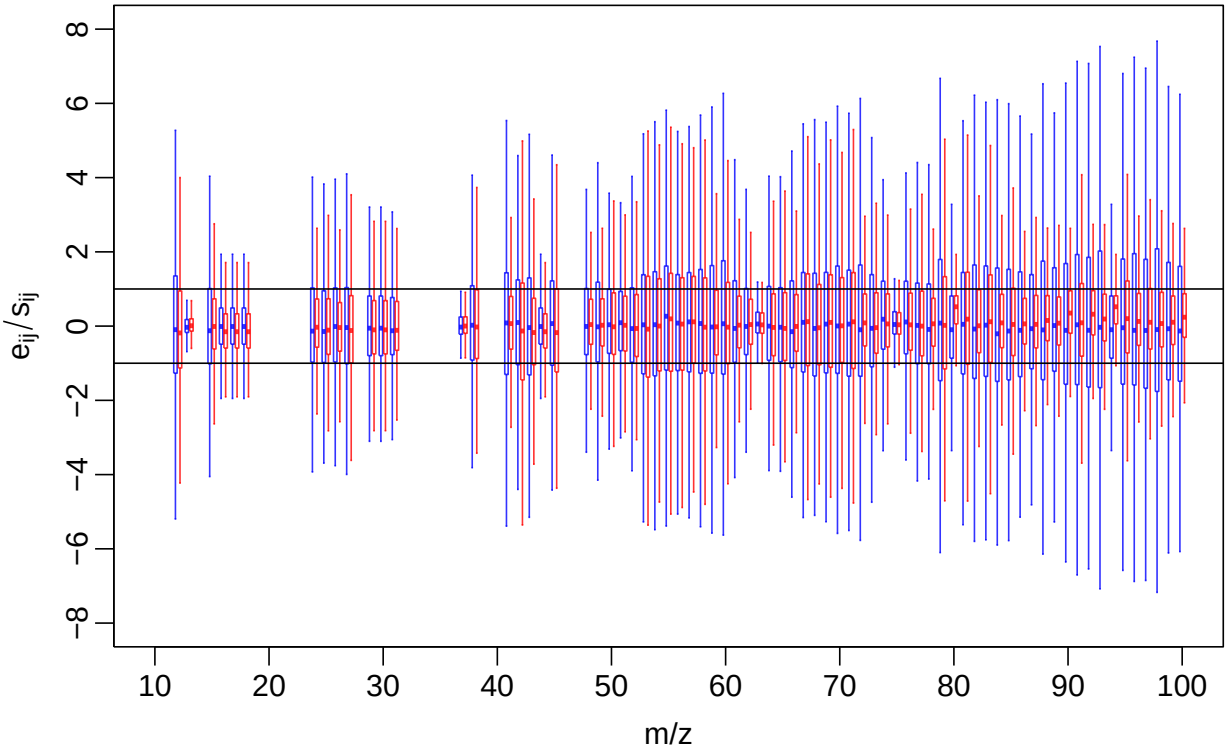

Figure S1: Residuals normalized by corresponding standard deviation values (original in blue; revised in red). Thick, horizontal lines indicate median values, boxes span interquartile range, and whiskers span 1.5 times the interquartile range. Black, horizontal lines are drawn at values of ± 1.

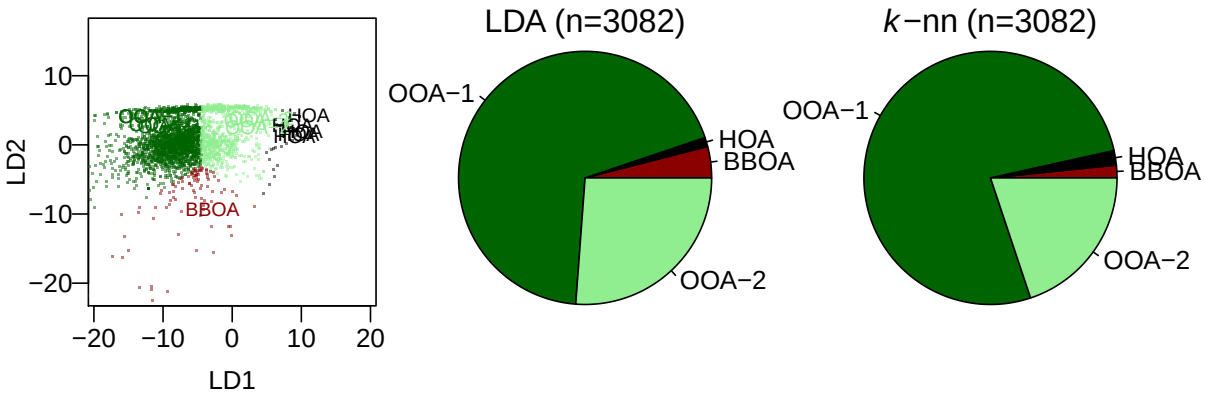

Figure S2: Feature vectors projected onto the space of the first two linear discriminants, and classified fractions of sample spectra according to LDA and k-nn methods (from left to right).

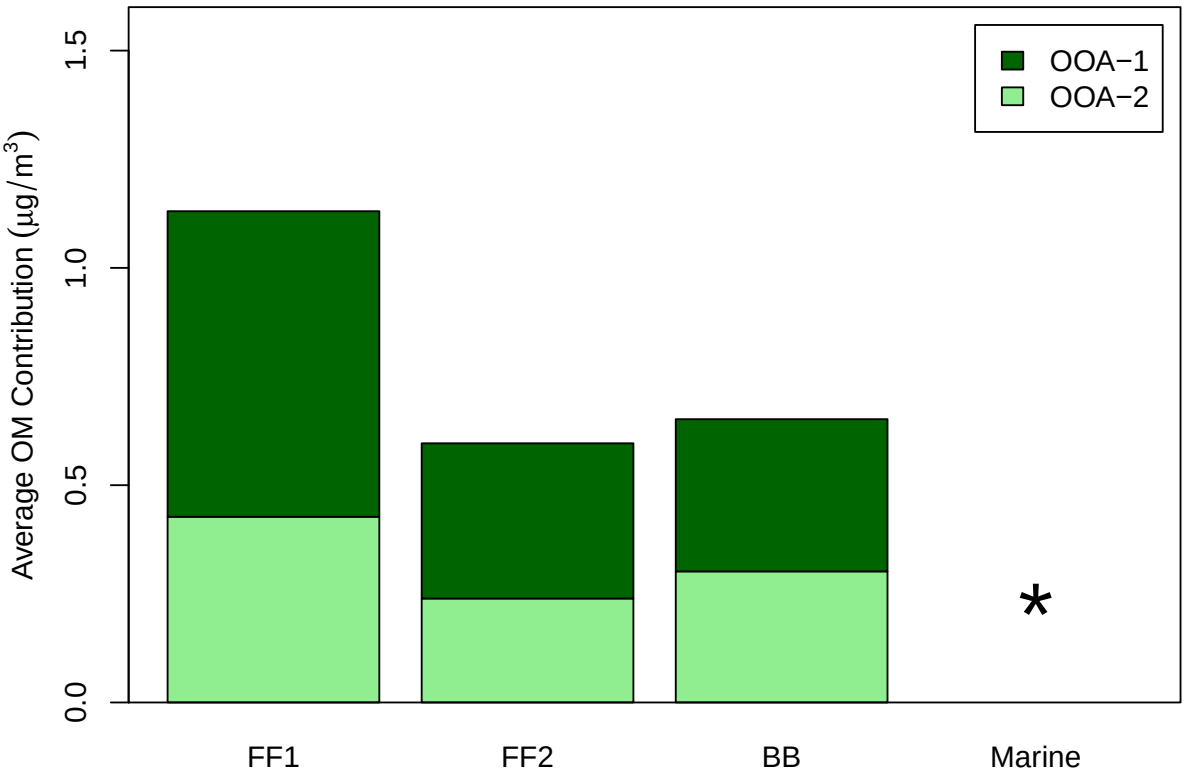

Figure S3: Proportion of FTIR PMF factors associated with ACSM PMF factors, estimated from Equation (S6). Asterisk indicates component not included in OM sum.

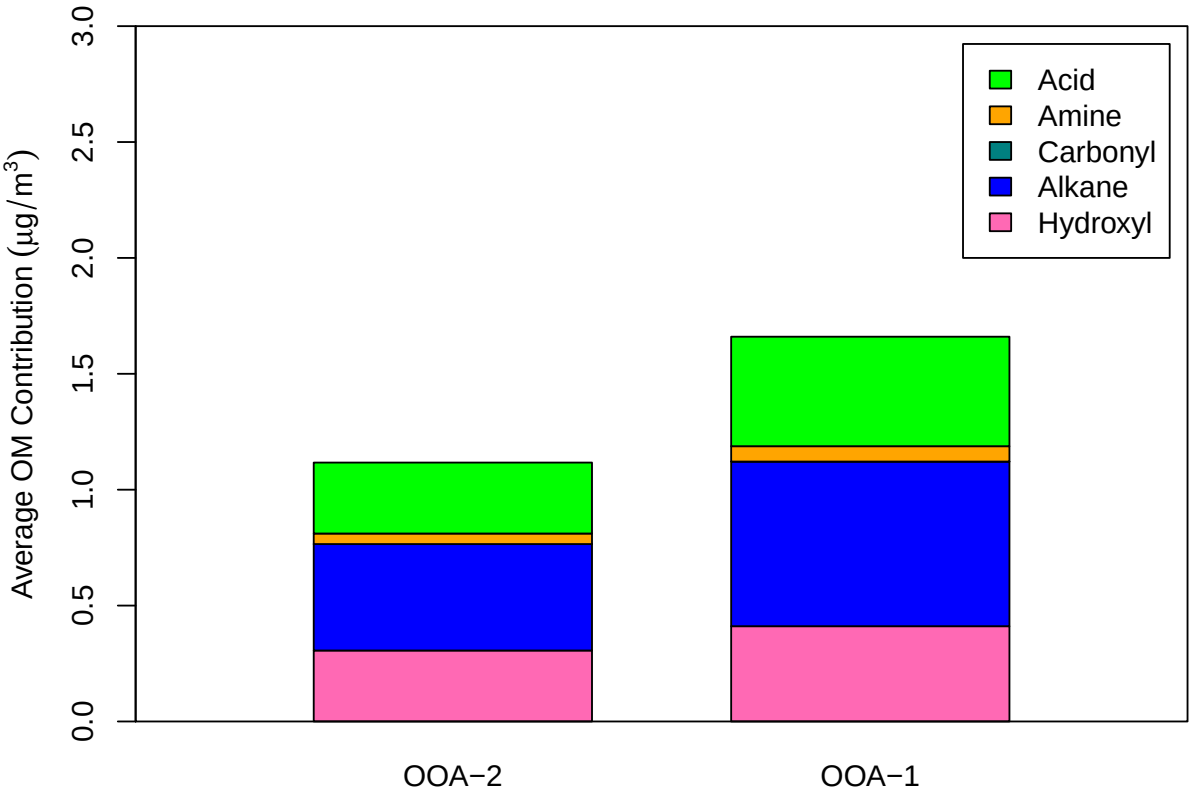

Figure S4: Proportion of FTIR (non-marine) OFG associated with ACSM PMF factors, estimated from Equation (S6) with modifications described in Section S2.1.

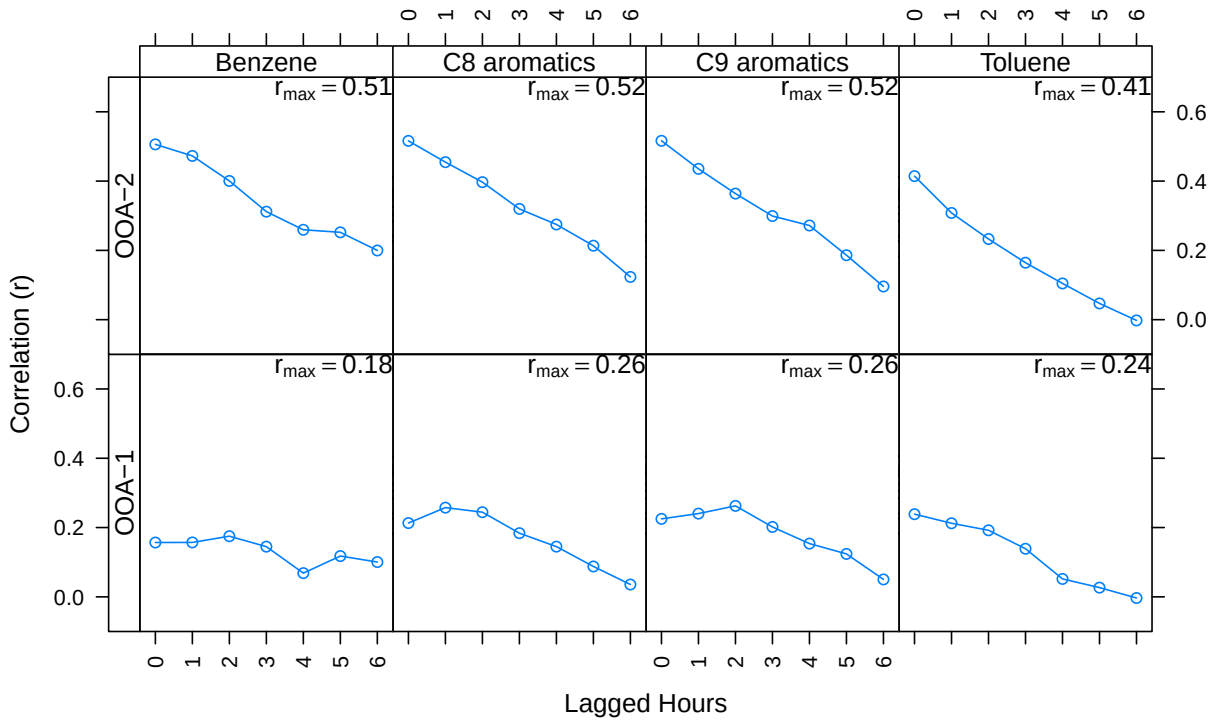

Figure S5: Lagged correlations of ACSM PMF factors with VOCs measured by PTR-MS.

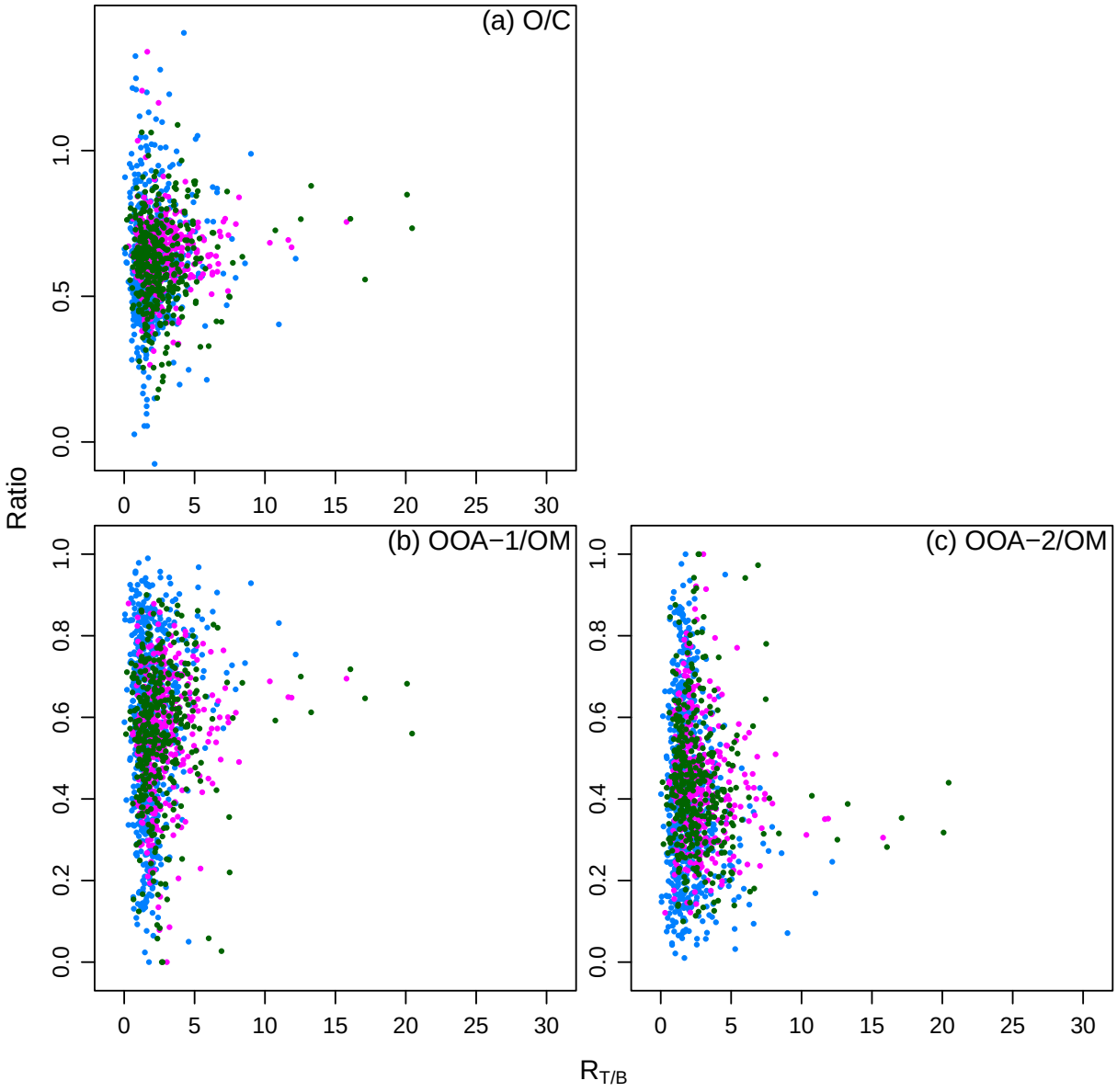

Figure S6: O/C ratio and PMF-factor fractions of OM as a function of Toluene-to-benzene ratio ($R_{[}T/B]$). Color indicates time of day: evening (blue), morning (pink), afternoon (green).

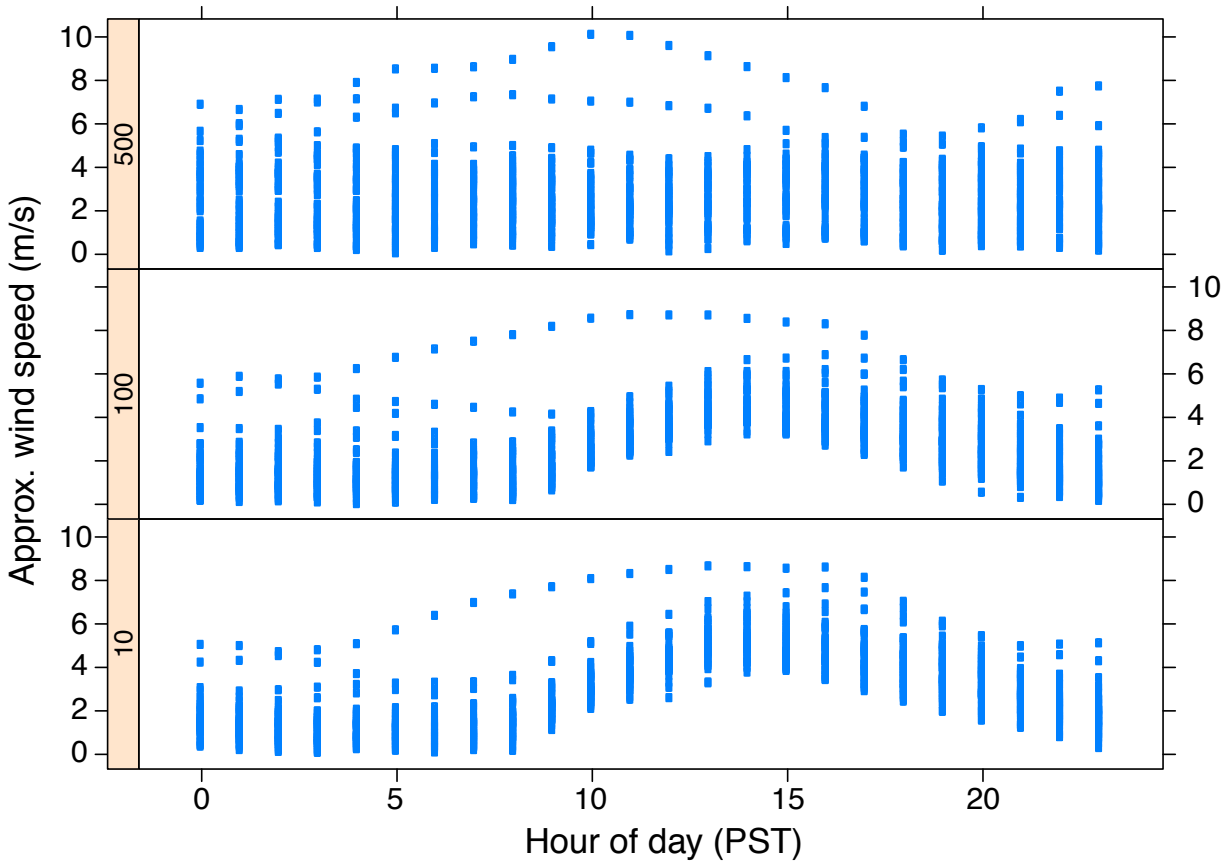

Figure S7: Wind speeds estimated by HYSPLIT model at altitudes of 10, 100, and 500 m shown in vertical panels from bottom to top, respectively.

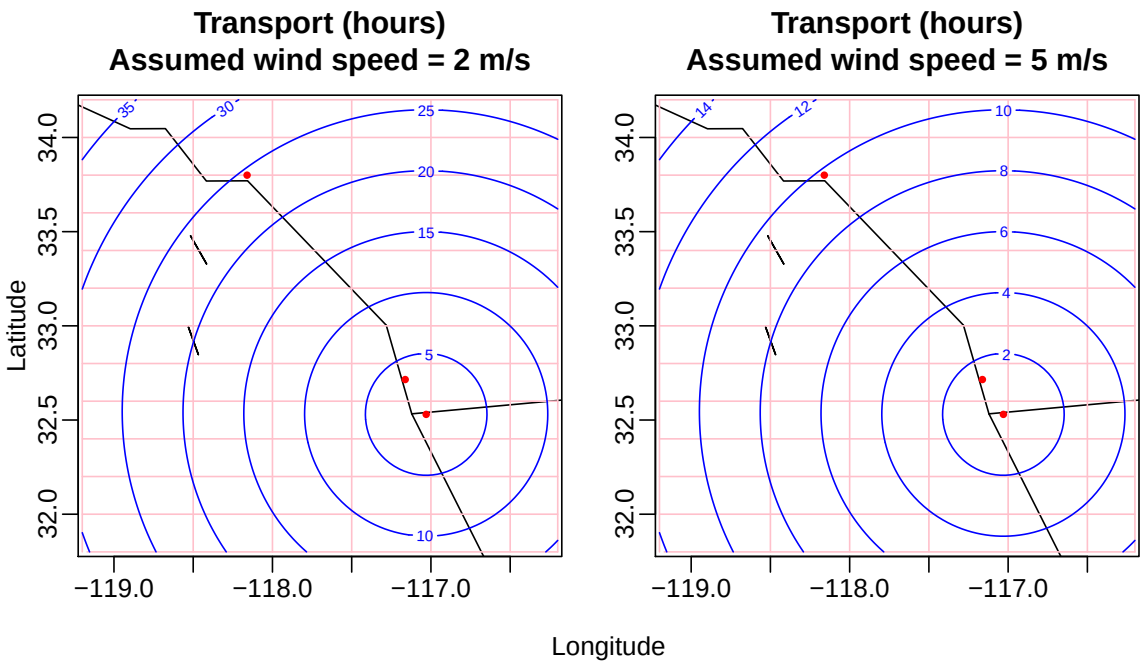

Figure S8: Transport times (blue) required for airmasses to reach Tijuana, estimated for fixed wind speeds of 2 and 5 m/s. Red dots from north to south represent Long Beach, CA, San Diego, CA, and Tijuana (Parque Morelos), Mexico.