Atmos. Chem. Phys. Discuss.,
https://doi.org/10.5194/acp-2018-427-RC1, 2018

Atmospheric Chemistry and Physics Discussions

Interactive comment on "An automatic observation-based typing method for EARLINET" *by* Nikolaos Papagiannopoulos et al.

S. P. Burton (Referee)

sharon.p.burton@nasa.gov

Received and published: 28 June 2018

General:

This paper describes the adaptation of a flexible automated typing algorithm to EARLINET network lidar data. It's great to see this product being produced for such a large and continuing dataset. As the authors point out, aerosol typing is potentially useful for developing a better understanding of aerosol sources and for improving the accuracy of satellite retrievals and climate and weather models. Producing typing data for EARLINET makes some of these goals become more accessible. The scientific methodology is good and the analysis and testing are thorough and include the introduction of useful new statistics and tools. The success rate of the algorithm is not

always high, but the authors present an analysis of how this depends on the observed variables and how the algorithm would be adapted for making use of additional variables. The precision of the language could be improved, and I have a few suggestions about this and about some other aspects of the analysis below.

Specific comments:

Figures 2 and 7. What are the resolution of the extinction and backscatter profiles? How is the lidar ratio calculated? Were the extinction and backscatter at the same resolution before taking the ratio? I ask because there are differences in the shape of extinction and backscatter in each of the discussed layers that do not seem particularly consistent with the idea that each layer is a specific coherent aerosol type. For example, the "layer" below 2 km in Figure 2 has a completely different shape in backscatter vs. extinction, leading to large variability in the lidar ratio, much larger than the suggested error bars. Do you think this variability is real, or could it be that the local maximum (seen in backscatter) is smoothed out in extinction by a coarser vertical resolution? If it's real, is it likely this is a single consistent aerosol type in this layer? Similarly, the different slopes in the "layer" between 3.5 and 5 km lead to a very large slope in the lidar ratio that does not seem consistent with the idea that this is a single aerosol type. If this variability is spurious, it is liable to create additional apparent noise in the classification that is not really related to aerosol variability within classes. (Of course, spurious error would also be a concern in general, not just for classification.)

Would you say that there is a possibility for error in the determination of "truth" aerosol types? If so, it would be good to see some discussion of that. I also have a particular question about the interpretation of the influence of marine aerosol in the two case studies that are discussed at length. FLEXPART, Figures 3 and 8, seems like a very nice tool for information on aerosol source. Figure 3 seems to show a large fraction of the incidence below 2 km (a lot of the green and yellow) as being in the Mediterranean Sea, but in the discussion, no mention is made of marine influence and the case is described as "pure dust". On the other hand, in Figure 8, an apparently lesser pro-

portion of the trajectories below 2 km are seen over the Black Sea and the Caspian Sea, and this layer is described as a mixture "enriched with marine particles during their overpass over the Black Sea". Can you clarify how we can know that there is marine influence in one case and not the other? It seems that it would be particularly difficult to say definitively when aerosol types are "pure" rather than mixed, using this method. Can you clarify whether there are additional factors that go into these judgements besides the FLEXPART tool and what uncertainties are associated with those judgements?

Section 3.2.4. When you add particle depolarization as a classification variable, I think you should still keep the original three variables. You have already shown that all three variables are sufficiently independent to be useful, so adding an independent fourth variable would be expected to produce the best classification method. I'm not following why you avoid using more than 3 variables.

P4 L14 The idea that the spectral ratio of lidar ratio indicates smoke aging is still a hypothesis, based empirically on a small number of suggestive cases. Describing this relationship as "robust" overstates the case, I think. Not all of the references actually support the statement. For example, Samaras et al. 2015 have no measurements related to smoke age, but rather take it as given, substituting the spectral ratio of lidar ratios as a proxy for smoke age. Please don't use a reference to support a hypothesis that merely made use of the hypothesis (at least not without more explanation). Anyway, the current manuscript doesn't relate to smoke aging. You could easily remove the statement and avoid controversy. At least don't overstate it and please remove references that do not really support the statement.

Should the "clean" in the label "clean continental" be taken literally? The description of the clean continental category is obliquely defined here as a mixture of polluted continental and clean marine aerosol and indeed the data in Figure 5 also seem to support its interpretation as a mixture of the two. I think this is an interesting way to think about this type, perhaps much more useful than the standard way of thinking

of it as a type that, unlike all the others, is defined by an extensive aerosol property (low aerosol loading). Any comment on this? Would a case that has the intensive properties of the clean continental class but a significant amount of aerosol optical depth (so therefore not particularly "clean") be considered "clean continental" in your analysis?

P9-10. The information about Mahalanobis distance is basically repeated from earlier work. You could simplify by referencing Burton et al. 2012 and noting the different thresholds.

P10, L7. Similar probabilities for two different classes do not indicate mixing between those two classes. This only reflects that those two classes are close to each other in your measurement space. For example, any point that is close to your "smoke" class is also close to your "polluted continental" class because the classes are close to each other. If it actually is smoke plus a little bit of marine influence, the 2nd closest class will still be "polluted continental", not marine.

Section 3.2.2 classifying variables selection. I think choosing variables solely based on Wilks' partial lambda may not catch everything. It may be that different variables have more power to separate different subsets of classes. For example, depolarization obviously has a lot of power to separate dust classes and almost no power to separate non-dust classes. If you had a variable that helped separate smoke from polluted continental, even partially, but did nothing else, it may have a poor partial lambda but it would nevertheless be extremely valuable. I suggest a plot similar to figure 10 in Burton et al. 2012 as a way to understand more thoroughly what each variable contributes to separating the classes. By looking at the variability of each variable within each class you can see where the overlaps are in every dimension. Your figure 5 also does this but it's usefulness maxes out at 2 dimensions, since it is very difficult to visualize more than 2 dimensions in a literal space. Your Wilks' lambda analysis suggests that the ratio of lidar ratios has significant discriminatory power, but Figure 5 does not reveal how (that is, whether some sets of classes that look like they overlap might be distinguished by

the 355 nm lidar ratio or the spectral lidar ratio). It would be nice to have a visualization that answers that question. This would also address the question of whether spectral lidar ratio separates smoke and pollution aerosol (Muller et al. 2007a) which would be an interesting discussion in itself.

P16, Perhaps you could discuss more explicitly the tradeoff between more classes and less classes. All of your statistics (except Wilks' lambda) seem to show better performance with fewer classes, but that could be taken to an extreme. That is, with only one class, there would be no errors at all! How do you address this tradeoff?

Do you plan to share these aerosol typing results publically? What about the training database of manually typed samples? Also please include links to the EARLINET database.

Typos and requests for clarification:

P2, last sentence: Both cluster analysis techniques and supervised classification techniques need the number of groups as input.

P4 L7: "operates" should be "operate"

P5, L18: talks about the region of incomplete overlap. What altitude does this go up to?

P5, L22-23: the sentence "the aforementioned layers" is unclear and should be reworded. I think you are suggesting that the intensive properties are approximately constant throughout each layer, but that does not really appear to be true, so perhaps I'm misunderstanding the wording.

P5, L24 and throughout: there are four Angstrom exponents discussed but often the text refers to "Angstrom exponent" as if there is only one. Please clarify which one you mean. Likewise, it should be specified which wavelength is meant when "lidar ratio" is used.

P6 L23. I recommend taking more care about using the word "absorbing". It seems that "more absorbing" is here used as a synonym for "higher lidar ratio", but lidar ratio depends on particle size and other factors as well as light absorption and is really not as direct an indicator as this language suggests. Also P11 L4 and L14 (continental pollution is not necessarily absorbing); P8 L5 (smoke is often absorbing but not always highly absorbing especially when aged); and perhaps elsewhere.

P7, L14. Delete "mainly". Although the variability in the lidar ratio is a "hot topic" there is also significant variability in, for instance, extinction Angstrom exponent.

P7, L31. The suggestion for CALIPSO to add a dust+marine type was made several times before 2016 also, for example Kim et al. 2013, Burton et al. 2013, Rogers et al. 2014

Kim, M.-H., Kim, S.-W., Yoon, S.-C., and Omar, A. H.: Comparison of aerosol optical depth between CALIOP and MODIS-Aqua for CALIOP aerosol subtypes over the ocean, Journal of Geophysical Research: Atmospheres, 10.1002/2013jd019527, 2013.

Rogers, R. R., Vaughan, M. A., Hostetler, C. A., Burton, S. P., Ferrare, R. A., Young, S. A., Hair, J. W., Obland, M. D., Harper, D. B., Cook, A. L., and Winker, D. M.: Looking Through the Haze: Evaluating the CALIPSO Level 2 Aerosol Layer Optical Depth using Airborne High Spectral Resolution Lidar Data, Atmos. Meas. Tech. , 7, 4317-4340, 10.5194/amt-7-4317-2014, 2014.

P8 L10 and in the references: I think Pereira should be Nepomucino Pereira (that is, the first author appears to have a two-part surname).

P9 first paragraph and third paragraph: There are a few places, including these 2 paragraphs, where the wording is awkward with several errors in English language usage that make them hard to understand. Please reword for clarity.

P12 L28, I think you mean Burton et al. 2015, not 2014. Burton, S. P., Hair, J. W.,

Kahnert, M., Ferrare, R. A., Hostetler, C. A., Cook, A. L., Harper, D. B., Berkoff, T. A., Seaman, S. T., Collins, J. E., Fenn, M. A., and Rogers, R. R.: Observations of the spectral dependence of linear particle depolarization ratio of aerosols using NASA Langley airborne High Spectral Resolution Lidar, Atmos. Chem. Phys., 15, 13453-13473, 10.5194/acp-15-13453-2015, 2015.

P13, L25 & L27, and throughout. When you count measurements or samples, what defines a single measurement, given that the lidar systems operate basically continuously? Please discuss how you select data and discuss what criteria are used in data selection and how much averaging is done.

P13, L25 & L27. Also, the numbers in this paragraph are confusing. If there are only 42 measurements available with three wavelengths, how are there 47 samples? Of if you don't need all 3 wavelengths, then why only 47 instead of 157?

P14-15, discussion of recall, precision and accuracy (also error rate from the earlier discussion). The description of these variables could be made clearer and it should be specified that recall and precision refer to specific classes. I think the use of the terms false negative and false positive contributes to confusion rather than clarity; the descriptions are too similar to distinguish them. I think you are saying that recall for a particular class indicates the number of correct identifications divided by the number of actual instances of that class. Precision for a particular class indicates the number of correct identifications of that class divided by the number of times when that class was predicted, whether rightly or wrongly. Am I understanding it right? And what is "accuracy"? Is it defined the same as "error rate" or does "accuracy" only count the instances that are classified and not the ones that are unclassified due to overlap between classes? Are all four of these statistics independently useful, or could it be simplified by using fewer?

P15 L9 "cannot be evaluated". It seems that even though they don't appear in the training set, they were occasionally predicted by the method. If so, perhaps reporting

how often that happened would be useful.

Table 1. I don't understand what the value in parentheses represents.

Table 2. Am I interpreting this correctly, that smoke is never predicted by the classification methodology? Any comments about implications of that?

Table 2. I think PC+C in the 7th column on the top should be PC + S, and I think the 2nd CC in the 4th column on the bottom should be S.

Figure 4. What does "trained classifier" in the middle of the flowchart mean? How is it different from "typing procedure"?

Interactive comment on Atmos. Chem. Phys. Discuss., https://doi.org/10.5194/acp-2018-427, 2018.