

Automatisierte Excel-Cluster-Analyse verifiziert an Rohwerten von Elementen im PM10-Staub eines Jahres an einer UBA-Messstelle

Wolfgang Grosch
bei der Abteilung Luft des Umweltbundesamtes
Außenstelle Langen
D 63225 Langen, Paul-Ehrlich-Straße 29

Zusammenfassung

Die Multikomponenten-Element-Analyse mit ICP (Induktiv gekoppeltes Plasma) und anschließender Massenspektroskopie ist fester Bestandteil der Umweltanalytik. Im Luftmessnetz des Umweltbundesamtes wird es u.a. zur Analyse von Feinstaubinhaltsstoffen im Rahmen des europäischen Messprogramms EMEP eingesetzt.

Eine vorgegebene Excel-Mappe, in die ausgewählte Messwerte der ICP mit einem Visual-Basic-Programm des Autors importiert werden, um dann mit bestehenden Verknüpfungen und Grafiken verarbeitet zu werden, wenn neue Messdaten importiert sind, wurde so ergänzt, dass zusätzlich online eine automatisierte Clusteranalyse zwischen den Staubinhaltsstoffen durchgeführt wird.

Die mit Excel dann sehr einfach von Hand komplettierte hierarchische Clusteranalyse für die an einer UBA-Messstelle in einem Jahr beobachteten Rohwerte der Feinstaubinhaltsstoffe ergab beispielhaft einen Cluster zwischen den beiden Eisenisotopen und einen weiteren Cluster für den zweiten eindeutigen Zusammenhang zwischen Blei, Tellur, Antimon, Cadmium und schwächer Zink und davon getrennt einen dritten Cluster für Mangan, Kobalt, Nickel.

Diese Cluster weisen auf gemeinsame Ausbreitungshistorie und Gleichgewichtszustände zwischen emittierten Gasen und Feinstäuben hin. Sie stehen offensichtlich am Schluss einer Kette von physikochemischen Prozessen, bilden das Ergebnis ab und lassen trotzdem noch Rückschlüsse auf gemeinsame Quellgruppen zu. Die anderen Inhaltsstoffe erleiden Individualschicksale. Das Verfahren wird in Zukunft zur Qualitätssicherung der Messungen eingesetzt und künftige Cluster-Analysen der Staubinhaltsstoffe für alle UBA-Messstellen ermöglichen.

Eine Möglichkeit und Notwendigkeit zur Beschränkung der Analyse auf Leitsubstanzen für die Cluster wird nicht gesehen.

1. Einleitung

Im Zuge der Automatisierung von Staubinhaltsstoffanalysen mit dem Verfahren der ICP wurde eine von Bieber (1) entwickelte Excel-Vorlagen-Mappe mit Verknüpfungen für den aktualisierenden Import von Textdateien mit Messwerten des Analyseautomaten so fortentwickelt, dass die jährlichen Wochen-Messwerte der Staubinhaltsstoffe einer teilautomatisierten Cluster-Analyse nach einem hierarchischen Prinzip unterzogen werden. Lediglich die hierarchische Prüfung der angezeigten primären Primärcluster-Paarungen muss von Hand mit Hilfe einer programmierten Excel-Funktion zur Errechnung der Euklidischen Distanz einer Messreihen-Paarung durchgeführt werden. Mit diesem Verfahren wurden zur Validierung beispielhaft die Rohwerte der PM10-Staubinhaltsstoffe von einer UBA-Messtelle einer Cluster-Analyse unterzogen.

2. Gang der Analyse

Zur Analyse werden die ursprünglich gemessenen Konzentrations-Messreihen in ng/m³ mit der gleichnamigen Excel-Funktion standardisiert auf den Mittelwert 0 mit der Streuung 1 und in Tabellen-Spalten abgelegt. Aus diesen standardisierten Messreihen-Spalten lassen sich dann nach dem Schema von (3) die Summen der Fehlerquadrate zwischen den standardisierten Reihen und damit die Euklidischen Distanzen für jede mögliche kombinatorische Paarung (alle möglichen Permutationen) der a,b-Inhaltsstoffe errechnen. Die Euklidische Distanz für a_n, b_n-Messwerte ist gegeben durch die Wurzel (dem Abstand des Paares im n-dimensionalen Raum) aus der Summe der Fehlerquadrate:

$$ED(a, b) = \sqrt{\sum_n (a_i - b_i)^2}$$

Zur Berechnung wird die Excel-Funktion (4) eingesetzt, die über „Extras, Makro, Makros, bearbeiten“ mit dem Editor in die Mappe eingegeben wird.

Das Kombinations-Tabellen-Dreieck der Euklidischen Distanzen wird aufsteigend sortiert (die Kombinations-Matrix ist symmetrisch aufgebaut, daher wird nur ein Dreieck berechnet). Die Euklidische Distanz 0 steht für völlige Übereinstimmung der Streuung der Messwerte des Reihenpaares. So ist ED (a,a) = 0 und ED (a,b) = ED (b,a), daher bleiben diese Trivillösungen außer Betracht. Größere Werte kennzeichnen mangelnde Übereinstimmung der zeitlichen

Streuungen um den Mittelwert. Das Paar hat in diesem Fall keine Koinzidenzen. Hat man ein Paar mit guter Übereinstimmung gefunden, so kann man mit Hilfe von Excel neue Reihen aus den Schwerpunktmesswerten einer Messreihen-Paarung nach (2) zu

$$ab_i = \frac{a_i + b_i}{2}$$

errechnen. Diese neuen Reihen lassen sich mit anderen Paarungs-Schwerpunkt-Reihen wieder mit Hilfe der Euklidischen Distanzen vergleichen. Ist die Distanz der Schwerpunkt-Paarungen kleiner als die der ursprünglichen kombinatorischen Paarungen, wird ein Zusammenhang zwischen den höherrangigen Schwerpunkt-Paarungen bestätigt. Daher wird dieses Verfahren der Erprobung von Schwerpunktreihen hierarchisch genannt. Für den Fall von Dreierclustern kann man die Schwerpunktreihe von drei Reihen nach (2) durch

$$abc_i = \frac{a_i + b_i + c_i}{3}$$

abschätzen. Das Verfahren der Prüfung eines Zusammenhangs zwischen zwei Schwerpunktreihen mit Hilfe der Euklidischen Distanz bleibt gleich.

3. Validierung der Analyse: Cluster im PM10-Staub

Mit Hilfe dieses oben beschriebenen Verfahrens wurden die Rohwerte der Jahres-Wochenwertreihen der Messwerte von Arsen, Cadmium, Kobalt, Kupfer, Eisen⁵⁴, Eisen⁵⁷, Mangan, Molybdän, Nickel, Blei, Antimon, Selen, Strontium, Tellur, Vanadium, Zink einer Clusteranalyse unterzogen.

Dabei ergab sich z.B. zwischen den beiden Eisenisotopen die kleinste Euklidische Distanz von 0.4. Diese bilden damit das erste Cluster mit bester Übereinstimmung. Mit Distanzen von 2-4 steht eine Reihe von Paarungen an zweiter Rangstelle, wie aus der Dreiecksmatrix in Abbildung 1 mit Euklidischen Distanzen von möglichen Kombinationen der Reihen ersichtlich ist.

Eine erste Sichtung auf mögliche Cluster lässt erwarten, dass die Paarungen bei Blei und bei Mangan die ersten Hinweise auf mögliche Cluster geben. Die möglichen Kombinationen wurden schwerpunktmäßig näher nach dem beschriebenen Muster untersucht. Dabei bestätigten sich diese ersten Erwartungen, wie die folgenden Ausführungen zeigen.

Abbildung 1:

Beispiele Euklidischer Distanzen für alle möglichen Staubinhaltsstoff-Paarungen für die UBA-Messstelle

	As	Cd	Co	Cu	Fe54	Fe57	Mn	Mo	Ni	Pb	Sb	Se	Sr	Tl	V	Zn
As		5.2	8.8	7.6	9.6	9.7	9.0	8.4	9.1	4.2	5.6	7.7	8.8	6.0	10.2	6.2
Cd			8.5	7.1	9.4	9.5	8.7	8.0	8.9	2.6	2.9	7.5	8.2	3.0	10.0	4.5
Co				5.6	6.1	6.2	3.2	9.1	4.4	9.1	8.9	10.8	8.4	9.4	8.5	8.9
Cu					8.7	8.8	4.2	6.5	4.4	7.2	7.4	8.8	7.8	7.8	10.3	6.9
Fe54						0.4	6.6	9.8	9.1	9.5	9.3	10.2	7.3	9.7	7.6	9.6
Fe57							6.6	9.8	9.1	9.6	9.4	10.3	7.4	9.8	7.6	9.7
Mn								8.6	3.5	9.2	9.2	10.4	8.5	9.6	9.6	8.6
Mo									8.3	7.8	7.1	6.6	9.8	8.5	10.2	7.3
Ni										9.5	9.4	10.5	10.0	9.8	9.8	8.7
Pb											3.1	7.1	7.4	2.8	9.9	4.3
Sb												6.7	8.4	3.8	10.0	4.5
Se													9.3	7.3	11.1	7.2
Sr														7.0	8.8	8.7
Tl															9.8	5.1
V																10.2
Zn																

Untersucht man die Schwerpunktreihen-Kombinationen von Blei+Cadmium, Antimon+Cadmium, Cadmium+Tellur, Blei+Tellur und auf der anderen Seite Mangan+Kobalt, Kobalt+Nickel und zieht noch die Dreier-Schwerpunkte Cadmium+Blei+Antimon und Mangan+Nickel+Kobalt hinzu, so erkennt man aus Abbildung 2 den zweiten getrennten Cluster Blei+Cadmium+Antimon+Tellur den man noch durch Zink ergänzen kann und dritterseits Mangan+Nickel+Kobalt als verifiziert:

Abbildung 2:

Beispiele Euklidischer Distanzen für Schwerpunkt-Kombinationen zwischen den Staubinhaltsstoffen aus Abbildung 1

	Pb+Cd	Pb+Tl	Sb+Cd		Cd+Pb+Sb
Pb+Cd		1.5	1.5	Pb+Tl	1.8
Pb+Tl			1.9		Cd+Pb+Sb
Sb+Cd				Mn+Ni+Co	8.6
	Ni+Co	Mn+Ni			
Ni+Co		1.6			
Mn+Ni					

kein Zusammenhang

Dieses Ergebnis bedeutet, dass die betreffenden Element-Gruppierungen miteinander nicht nur bei der Immission vergesellschaftet sind. Das Auf- und Ab in

den Messreihen dieser Gruppen verläuft synchron, so dass sie physikochemisch gesehen einen vergleichbaren Gleichgewichtszustand in der Atmosphäre erreicht haben und nur noch Verdünnung, Konglomeration und Deposition unterliegen. Die Hypothese I lautet daher: Die Stoffe in den Clustern stammen aus den gleichen Quellgebieten und haben sich miteinander ausgebreitet.

Mögliche Quellen sind anthropogene Hochtemperatur-Verbrennungsprozesse für die Elemente im großen Cluster. Allerdings gilt Antimon auch als typisch für Bremsenabrieb, in dem es mit Anteilen von 5% vorkommt, so dass auch der Verkehr als anteilige Quelle dieser Gruppe in Betracht zu ziehen ist (5). Alle diese Elemente werden beim atmosphärischen Transport deutlich angereichert und bleiben nicht zwangsweise im teilweise gasförmigen emittierten Aggregatzustand (6), unterliegen aber offensichtlich einem relativ schnellen vergleichbaren physikochemischen Umsetzungs- und Ausbreitungsverhalten innerhalb weniger Stunden bis Tage, anders sind die starken Korrelationen innerhalb der Wochenwert-Cluster nicht zu erklären. Das würde für die Hypothese I sprechen, dass die Quellgebiete im weiträumigen Immissionsgeschehen noch abgebildet werden.

Andererseits vertreten Pleßow und Heinrichs (5) die Meinung: „Im dicht besiedelten und stark industrialisierten Mitteleuropa ist eine Vielzahl weiträumig verstreuter, anthropogener Emissionsquellen vorhanden, deren Emissionsprofile im Mittel offenbar eine gleichförmig zusammengesetzte Aerosolkomponente ergeben, ... die sich sowohl in der Luft als auch im Boden industrie- und verkehrsferner Gebiete wieder(findet).“ Trifft dies auf Wochenmittel zu, würde ein Rückschluss auf Quellgebiete aus der Clusteranalyse unmöglich, dies stellt also Hypothese II dar. Ohne Emissionskataster und Modellrechnungen für diese Quellen würden sich keine Zusammenhänge zwischen wöchentlicher Emission und Immission herstellen lassen.

Was spricht nun gegen eine solche Hypothese II? Der relative Verlauf der Immissionszeitreihen innerhalb der Cluster ist identisch, was einen Transport in der gleichen Luftströmung nahelegt, also auch die Herkunft aus einem bestimmten Quellgebiet. Innerhalb des Quellgebiets und bei der Ausbreitung kann es durchaus zu vergleichbaren physikochemischen Prozessen dieser Clusterelemente gekommen sein, die zu einer Gleichgewichtszusammensetzung geführt haben. Träfe Hypothese II zu, müsste es abhängig von den Ausbreitungsbedingungen zu einem einheitlichen Cluster von Elementen kommen, die nicht nach Zeitreihenverläufen aufsplittbar wären. Das könnte für die Jahresmittelwerte der Immission zutreffen, was das „ im

Mittel“ des obigen Zitats wohl eher meint. „Im Mittel“ könnte aber tatsächlich noch längere Zeiträume umfassen. Der Timescale ist hier wesentlich.

Ein Zusammenhang zwischen den Einzelsubstanzen in den Clustern besteht nur zwischen den Mitgliedern eines Clusters, nicht zwischen und mit den restlichen Clustern und Elementen. Wir sehen daher Hypothese I als erhärtet an, nach der ein Rückschluss auf Quellgebiete möglich ist. Nähere Rückschlüsse über mögliche Quellgebiete muss die künftige Analyse der qualitätsgeprüften Analysenwerte in Zusammenarbeit mit Emissionsexperten ergeben.

Da die ICP ohnehin eine Vielzahl von Elementen in einem Arbeitsgang untersucht, ist es sinnvoll, die Palette der derzeit analysierten Elemente auch in Zukunft nicht durch Auswahl von Leitsubstanzen einzuschränken, da das auch mögliche Minderungsmaßnahmen mit einer eingeschränkten Zielvorgabe belasten würde.

Literatur:

1. Elke Bieber, Umweltbundesamt II 4.5 (Luftmessnetz), Außenstelle Langen: ExcelMappe.xls per Email an den Autor
2. Psychologische Methodenlehre für Nebenfachstudierende von Dr. Andreas Wilm, www.psychologie.uni-kiel.de/methoden_nf/Inhalt/Skript/7.2_Clusteranalyse.pdf
3. Clusteranalyse mit Excel nach einer der hierarchischen Methoden (Single-Linkage) am Beispiel einer Sozialraumtypisierung der Stadt Ingolstadt, Helmut Schels, Vortrag in der AG-Methodik des Verbandes der Deutschen Städtestatistiker, www.muenchen.de/cms/prod2/mde/de/rubriken/Rathaus/40_dir/statistik/meldungen/vdst_2008/02_02_01_schels.pdf (Die Veröffentlichung enthält eine fehlerhafte Formel für die Euklidische Distanz)
4. Quelltext des Autors für eine Excel-Funktion zur Berechnung der Euklidischen Distanz für beliebig lange Wertereihen, die in Excel-Tabellen in Spalten nebeneinander angeordnet sind:

Option Base 1

```
Function EuklidDist(spalte1 As Range, spalte2 As Range) As Single
'Hinweis: die Funktion wird in der Befehlszeile mit einem die Bereiche
'trennenden ; aufgerufen, die standardisierten Wertespalten dürfen
'keine Leerzellen enthalten (Excel 2003)
```

```
Dim quadrat, zeile() As Single
Dim zelle As Range
Dim i, j, intz As Integer
intz = spalte1.End(xlDown).Row 'Endzeile Spalte feststellen als intz
ReDim zeile(intz) 'Dimension kann erst zur Laufzeit zugeteilt werden
```

```
For Each zelle In spalte1
i = i + 1
zeile(i) = zelle.Value
'beginnend bei zeile(1) Zellenwerte der spalte1 zwischenspeichern
Next zelle
j = 0
```

```
For Each zelle In spalte2
'eigentliche Abarbeitung der Differenzen zwischen den gleichzeitigen
'Zellen in spalte1 und spalte2
j = j + 1
quadrat = quadrat + (zeile(j) - zelle.value) ^ 2
'Für alle Zellen der Spalte Summe der Quadrate aus den quadrierten
'Differenzen bilden
Next zelle

EuklidDist = Sqr(quadrat) 'Quadratwurzel aus der Summe der Quadrate

End Function
```

5. Pleßow K, Heinrichs H (2000) Anthropogene Spurenelemente in Aerosolen industrie- und verkehrsferner Gebiete: 205-223 In: M. Huch, H. Geldmacher (Hrsg) GUG-Schriftenreihe "Geowissenschaften und Umwelt" Umweltgeochemie in Wasser, Boden und Luft. Geogener Hintergrund anthropogene Einflüsse, Springer-Verlag, Berlin
6. Graedel ,TE, Crutzen, PJ (1994), Chemie der Atmosphäre – Bedeutung für Klima und Umwelt, SpektrumAkademischer Verlag, Heidelberg Berlin Oxford: 511S.