

# Methodische Aspekte bei der Modellierung des Zusammenhanges zwischen einer Exposition gegenüber Bioaerosolen und respiratorischen Symptomen

Methodological aspects in modeling the relationship between an exposition to  
bioaerosols and respiratory symptoms

*Michael Hoopmann*

## **Abstract**

Taking data from a cross-sectional study we discuss the question whether any relationship between bioaerosols coming from livestock and allergic or respiratory symptoms of 5–6 year old children exist as an example of mixed exposures. Bioaerosols coming from livestock are a compound of different agents. They are highly correlated concerning their emission but may have different health-related outcomes. We used the first principal component instead of the original measures of exposure within a multivariate logistic regression model to adjust for multicollinearity. We have chosen a natural cubic spline approach to approximate arbitrary functional relationships between this new constructed exposure variable and different symptoms. These statistical techniques allow a flexible and unbiased model approximation. The results are difficult to interpret in terms of a potential regulation, because the effects cannot clearly assigned to one of agents in the aerosol.

## **Zusammenfassung**

Anhand einer Querschnittstudie zur Frage, ob es einen Zusammenhang zwischen einer Exposition gegenüber Bioaerosolen und respiratorischen oder allergischen Symptomen bei Einschülern gibt, wird das Problem Mischexposition illustriert. Aus Tierställen emittierte Bioaerosole setzen sich aus diversen Stoffen zusammen, deren Konzentrationen untereinander hoch korrelieren aber unterschiedliche gesundheitsbezogene Wirkmechanismen haben können. Zur Kontrolle der Multikollinearität wurde anstelle der einzelnen Expositionsvariablen deren erste Hauptkomponente in logistischen Regressionsmodellen eingebaut. Ein natürlicher Splineansatz wurde zudem gewählt, um beliebige funktionale Zusammenhänge dieser neuen Expositionsvariablen und verschiedenen Symptomen zu approximieren. Diese statistischen Verfahren erlauben eine flexible und unverzerrte Modellanpassung. Die Ergebnisse sind jedoch für eine mögliche Regulation schwer zu interpretieren, da Effekte nicht eindeutig einem Wirkstoff zugeordnet werden können.

## **Forschungsprogramm zum Problem Bioaerosole in Niedersachsen**

Luftgetragene Stäube und Aerosole aus Tierställen entstammen zum überwiegenden Teil organischem Material, wie Futtermitteln, Einstreu, Tierepithelien oder -federn und Tierexkrementen. Sie enthalten Mikroorganismen mit deren Endo- und Exotoxinen sowie Parasiten. Die Stäube absorbieren darüber hinaus auch gasförmige Stallluftbestandteile, wie Geruchsstoffe, Ammoniak und Desinfektionsmittel (Millner 2009). Bei beruflich exponierten Tierhaltern können Stallstäube in hohen Konzentrationen zu chronisch obstruktiven Atemwegserkrankungen und Allergien führen (Nowak 1998; Iversen 1999; Douwes 2003).

Seit Jahren wird gerade in Niedersachsen mit seinen nutztierintensiven Flächen daher diskutiert, ob aus der Arbeitsmedizin bekannte gesundheitliche Effekte von Stallstäuben auch bei der Wohnbevölkerung in der Nachbarschaft von Tierställen auftreten (Schlaud 1998). Das Land Niedersachsen hat deshalb in den Jahren 2000 bis 2005 das Untersuchungsprogramm „Gesundheitliche Bewertung von Bioaerosolen aus der Intensivtierhaltung“ aufgelegt (Hoopmann 2005). Dieses umfasste auch zwei umweltepidemiologische Querschnittstudien, mit denen der Frage nachgegangen wurde, ob es bei Kindern (AABEL-Projekt: Hoopmann 2006)

beziehungsweise Erwachsenen (NiLS-Projekt: Radon 2002; Radon 2005b), die in der Nachbarschaft zu Tierstallungen wohnen, zu Gesundheitsbeeinträchtigungen kommt. AABEL wurde – an die Schuleingangsuntersuchungen gekoppelt – flächendeckend in vier Landkreisen in der Region Weser-Ems durchgeführt, NiLS in vier Gemeinden dieser Region mit besonders hoher Viehdichte. Für beide Studien wurde eine identische Expositionsquantifizierung anhand von Ausbreitungsrechnungen für tierstallbezogene Bioaerosole durchgeführt.

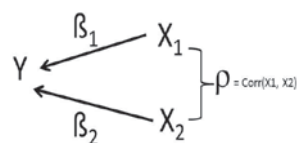
Abgesehen von grundsätzlichen Schwierigkeiten bei der Expositionsquantifizierung ergeben sich bei der epidemiologischen Auswertung besondere Probleme: Zum einen gab es praktisch keine auf die Anwohnerschaft abzielenden epidemiologischen Studien, die es erlaubt hätten, mögliche funktionale Beziehungen zwischen den Bestandteilen der Bioaerosole sowie der Wahrscheinlichkeit des Auftretens einer der Zielsymptome anzunehmen (O'Connor 2010). Insbesondere fehlten auch Angaben zu Wirkschwellen. Zum anderen korrelieren die einzelnen Bioaerosolbestandteile in ihrer Emission sehr stark miteinander, sodass man einen beobachteten Effekt beziehungsweise eine Assoziation zu den Symptomen kaum einem der Bestandteile eindeutig zuordnen kann.

Im Folgenden werden zur Veranschaulichung der Problematik Mischexposition exemplarisch Daten aus dem AABEL-Projekt vorgestellt. Für detailliertere Ergebnisse zu den Projektfragestellungen sei auf die jeweiligen Abschlussberichte und Originalpublikationen verwiesen (NLGA 2004; Hoopmann 2006; Radon 2005a; Radon 2005b).

## „Multikollinearität“ – lediglich ein statistisches Schätzproblem?

Regressionsmodelle modellieren simultan den Zusammenhang von mehreren Einflussgrößen (Regressoren) auf die interessierende Zielgröße, um so indirekte Effekte zu kontrollieren, die auf die Korrelation der Einflussgrößen untereinander zurückzuführen sind. Sobald die Regressoren jedoch zu stark korrelieren, kann die Schätzung der Regressionskoeffizienten aufgrund einer problematischen Matrixinversion extrem instabil werden. In diesem Falle spricht man von Multikollinearität, die besagt, dass einer der Regressoren nahezu als Linearkombination der übrigen Regressoren dargestellt werden kann.

Der einfachste Fall mit zwei korrelierenden Regressoren, die aber beide über einen anderen Wirkungspfad auf die Zielgröße wirken, illustriert das Problem:



In einem simulierten Beispiel mit 101 Beobachtungen, bei denen  $X_1$  und  $X_2$  extrem hoch korrelierten, wurden vier Modelle betrachtet.

Zunächst das übliche Regressionsmodell mit zwei Regressoren:

$$M_1 \text{ (wahres Modell): } Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

Die wahren Werte für die Simulation waren dabei  $\alpha = 0$ ,  $\beta_1 = \beta_2 = 1$ .

Als zweites ein unterspezifiziertes Modell mit nur einem der beiden Regressoren:

$$M_2 \text{ (unterspezifiziertes Modell): } Y = \alpha^* + \beta_1^* X_1 + \varepsilon^*$$

Als dritte Modellgleichung wurde eine Reparametrisierung des wahren Modells vorgenommen:

$$M_3 \text{ (reparametrisiertes Modell): } Y = \alpha + \gamma_1 X_1 + \gamma_2 (X_2 - X_1) + \varepsilon \text{ (somit } \gamma_1 = 2, \gamma_2 = 0)$$

In Erweiterung der bisherigen Betrachtung könnten die beiden Regressoren selbst fehlerbehaftet gemessen worden sein, was zur vierten Gleichung führt:

$$M_4 \text{ (Messfehler-Modell): } Y = \alpha + \beta_1 X_1^* + \beta_2 X_2^* + \varepsilon \quad X_1^* = X_1 + \zeta_1, X_2^* = X_2 + \zeta_2$$

Die Simulationsergebnisse sind in **Tabelle 1** wiedergegebenen.

Beim Vergleich der beiden wahren Modelle  $M_1$  und  $M_3$  erkennt man, dass zwar definitionsgemäß beide Modelle identisch anpassen (s.  $R^2$ ), der geschätzte Regressionskoeffizient für  $X_1$  unter  $M_1$  aber völlig unbefriedigend ist.  $M_2$  und  $M_3$  liefern beide als Parameterschätzung für  $\beta_1$  beziehungsweise  $\gamma_1$  den Effektschätzer für beide Regressoren zusammen. Während dies bei  $M_3$  in der Reparametrisierung begründet ist, ist bei  $M_2$  nicht bekannt, dass dieser Effektschätzer, der gemäß der Modellierung allein

Modell	R <sup>2</sup>	Schätzung für $\beta_1/\gamma_1$	entsprechender p-Wert	95%-Konfidenzintervall
M <sub>1</sub> : $Y = \beta_1 X_1 + \beta_2 X_2$	0,8565	0,698	0,193	[-0,359; 1,754]
M <sub>2</sub> : $Y = \beta_1 X_1$	0,8478	1,979	< 0,001	[1,811; 2,146]
M <sub>3</sub> : $Y = \gamma_1 X_1 + \gamma_2 (X_2 - X_1)$	0,8565	1,984	< 0,001	[1,821; 2,148]
M <sub>4</sub> : $Y = \beta_1 X_1^* + \beta_2 X_2^*$	0,8563	1,037	< 0,001	[0,566; 1,508]

auf  $X_1$  abzielt, auch  $X_2$  abdeckt.  $M_4$  liefert gerade im Vergleich zum entsprechend parametrisierten wahren Modell den besseren Schätzer für  $\beta_1$ . Kleine zusätzliche Fehler in den Regressoren führen somit zu stabileren Schätzern.

Bereits mit diesem einfachen Beispiel werden die Konsequenzen der Modellbildung unter Multikollinearität illustriert: Ein an sich korrekt spezifiziertes Modell kann zu instabilen Effektschätzern und damit zur Fehlinterpretation der Einzeleffekte führen. Unterparametrisierte Modelle hingegen sind zwar als Prognosemodelle geeignet, liefern jedoch verzerrte Schätzungen des Effektes eines Einzelregressors.

Die Modelle  $M_2$  bis  $M_4$  können als Prototypen für alternative statistische Verfahren (Sundberg 2002) angesehen werden, wie mit dem Problem der Multikollinearität beim wahren Modell umgegangen wird:

- Man modelliert nur einen Teil der Expositionsvariablen.
- Anstelle der Originalvariablen modelliert man mit (optimalen) Linearkombinationen.
- Man führt kleine Fehlerterme ein, die im Endeffekt die Schätzung stabilisieren.

Hierbei kann die Zielsetzung des Regressionsmodells entscheidend sein: Sollen Beobachtungen prognostiziert werden, interessiert man sich für eine möglichst gute Datenanpassung oder aber soll speziell der Effekt eines Regressors getestet werden?

## Lösungsansatz für das Multikollinearitätsproblem bei Bioaerosolen

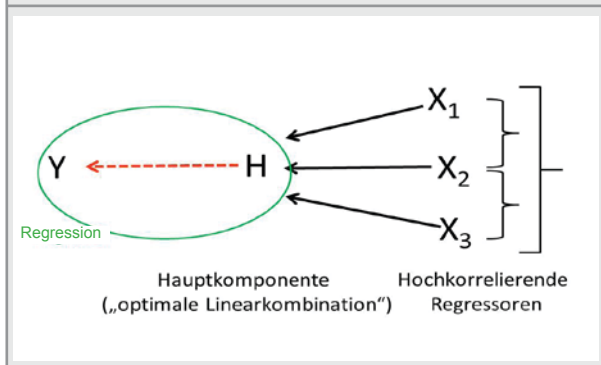
Die aus den Ausbreitungsrechnungen geschätzten Bioaerosol-Bestandteile korrelieren sehr hoch miteinander (**Tabelle 2**). Daher war bei der epidemiologischen Modellbildung zu überlegen, wie mit diesem Problem umgegangen werden sollte.

Im AABEL-Projekt wurde als Lösungsansatz allein die erste Hauptkomponente anstelle der vier Originalvariablen betrachtet: Bei der Hauptkomponentenanalyse (Hartung 1984; Sundberg 2002) wird eine Menge korrelierter Variablen in eine Menge unkorrelierter Variablen, die Hauptkomponenten, überführt (orthogonale Transformation). Jede einzelne Hauptkomponente ist dabei eine Linearkombination der Originalvariablen. Dabei werden die Hauptkomponenten in absteigender Bedeutung konstruiert, beginnend mit der ersten Hauptkomponente, die quasi die optimale lineare Kombination der Eingangsvariablen ist, die am meisten von der gemeinsamen Variabilität erklärt. Falls die Regressoren untereinander in etwa gleich stark korrelieren, entspricht die erste Hauptkomponente ungefähr dem Mittelwert der standardisierten Regressoren.

Würde man alle  $k$  Hauptkomponenten anstelle der ursprünglichen  $k$  Originalvariablen in einem Modell verwenden, so würde eine identische Datenanpassung erfolgen, da – wie im vereinfachten Eingangsbeispiel – die identische Information, nur linear transformiert, verwendet wird. Allerdings

	Pilzkonzentration	Gesamtkeime	Endotoxin, alveolengängig	Staub, alveolengängig
Pilzkonzentration	1	0,793	0,883	0,945
Gesamtkeime		1	0,960	0,960
Endotoxin, alveolengängig			1	0,986
Staub, alveolengängig				1

**Abbildung 1: Schematische Darstellung: Erste Hauptkomponente als Ersatz für hochkorrelierende Expositionsvariablen.**

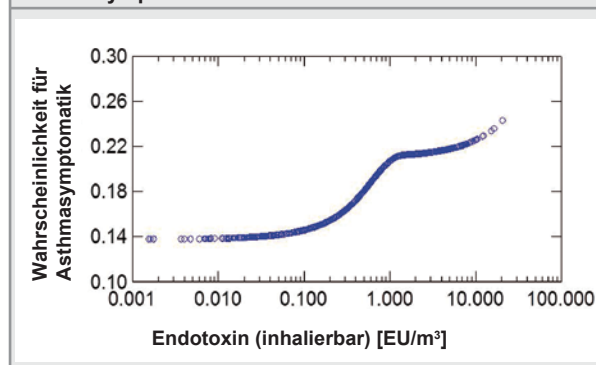


geht bei hochkorrelierenden Variablen kaum Information verloren, wenn man die niederrangigen Hauptkomponenten, die wenig zusätzliche Information beisteuern, nicht berücksichtigt. Indem anstelle der Originalvariablen nur die ersten Hauptkomponenten in das Modell aufgenommen werden, reduziert sich somit ohne großen Informationsverlust die Zahl der zu schätzenden Parameter.

Im AABEL-Projekt wurde allein die erste Hauptkomponente verwendet. Die erste Hauptkomponente als Form der Datenreduktion bedingt einen geringeren Informationsverlust als lediglich irgendeinen der Regressoren auszuwählen (**Abbildung 1**).

Neben der Multikollinearität besteht ein zweites Problem darin, dass die Form eines möglichen Zusammenhangs zwischen Bioaerosolen und den gesundheitlichen Zielgrößen unbekannt ist. Um die funktionale Beziehung zu approximieren, sind grundsätzlich verschiedene Ansätze möglich: Geht man von einem strikt monotonen Zusammenhang aus, können einfache lineare oder log-lineare Verläufe als Annahme durchaus begründet sein, bei unbekanntem Verlauf wählt man hingegen in der Epidemiologie häufig eine stufenförmige Approximation, bei der für die einzelnen Stufen, für die jeweils ein konstanter Verlauf angenommen wird, die Quantile der Expositionsverteilung herangezogen werden. Im AABEL-Projekt wurde demgegenüber für die Hauptfragestellung ein natürlicher Splineansatz gewählt, wobei Glättungsverfahren und insbesondere Spline gerade in umweltepidemiologischen Studien zunehmend Verbreitung finden (WHO 2004; Baksharan 2013). Dabei werden kubische Kurvenstücke an vorher festzulegenden Knotenpunkten miteinander verknüpft (Harrell 2001). Die Zahl der zu schätzenden Parameter bei einem natür-

**Abbildung 2: Beispiel für einen dreiknotigen natürlichen Spline für den Zusammenhang zwischen Endotoxin und Asthmasymptomatik.**



lichen Spline entspricht der Knotenanzahl minus 1. Wird somit anstelle der linearen Modellierung der Exposition mit einem Freiheitsgrad ein natürlicher Spline mit drei Knotenpunkten bei den Quartilen in ein (logistisches) Regressionsmodell eingebaut, so werden lediglich zwei Freiheitsgrade bei der Schätzung verwendet (**Abbildung 2**).

## Vorgehen beim AABEL-Projekt

Die Hauptfragestellung lautete „Gibt es einen gesundheitlichen Effekt der Exposition gegenüber Bioaerosolen aus Tierställen?“ Um diese Frage inferenzstatistisch zu bewerten, wurde wie folgt vorgegangen:

- 1) Die erste Hauptkomponente der vier Original-expositionsvariablen Endotoxin, Staub (jeweils alveolengängig), Gesamtkeimbelastung und Pilzkonzentration wurde gebildet.
- 2) Diese wurde als Regressor in einem logistischen Regressionsmodell für verschiedene a priori festgelegte allergische oder respiratorische Symptome herangezogen; somit fand eine Dimensionsreduktion statt, da anstelle der vier Originalvariablen nur eine Expositionsgröße verwendet wurde.
- 3) Die Expositionsgröße wurde nicht linear in das logistische Modell eingebaut, sondern ein Spline-Ansatz gewählt.
- 4) Die Festlegung der zusätzlich im Modell modellierten potentiellen Confounder basierte auf inhaltlichen Überlegungen, nicht jedoch im Zuge einer Variablenselektion, die zu verzerrten p-Werten führen würde.

5) Zusätzlich wurde für einige allergische Zielgrößen der Atopiestatus der Eltern als Effektmodifizierer behandelt.

Dieses Vorgehen führte zu unverzerrten p-Werten bei insgesamt 17 Endmodellen für die Zielsymptome. Von diesen führten zwei Einzeltests zu einem p-Wert von unter 0,05, drei weitere zu p-Werten unter 0,1. So zeigten sich asthmatische Symptome bei Kindern atopischer Eltern vermehrt, bei denen nicht-atopischer Eltern hingegen als verringert in Abhängigkeit einer Exposition gegenüber Bioaerosolen aus Tierställen.

Ergänzend zu diesem Vorgehen für die Hauptfragestellung wurde explorativ mit alternativen Expositionsparametern, einer eingeschränkten Studienpopulation oder auch alternativen Zielsymptomen gerechnet.

In der statistischen Endbewertung erwiesen sich wenige der hypothetischen Zusammenhänge als signifikant, sodass man – auch mit Blick auf das Studiendesign – nur von Hinweisen auf mögliche gesundheitliche Beeinträchtigungen sprechen kann. Allerdings erwiesen sich diese Hinweise über verschiedene Indikatoren als konsistent. Da die Expositionsquantifizierung am Wohnort die wahre Exposition nur approximieren kann, dürften die wahren Effekte eher systematisch unterschätzt werden.

## Diskussion des Vorgehens

Ein häufig berechtigter Kritikpunkt an epidemiologischen Modellen beziehungsweise am gesamten Auswertungsprozess besteht darin, dass der Modellbuilding-Prozess zu verzerrten Effektschätzern und p-Werten führt. Zudem werden Effektschätzer aus ihrem multivariaten Kontext herausgerissen und singular diskutiert, obgleich sie an sich nur unter den modellierten Bedingungen gelten. Bei AABEL wurde demgegenüber eine weitgehende Trennung der a priori formulierten Modelle und der anschließenden Exploration vollzogen. Zwar fand keine Kontrolle für multiple Vergleiche im Sinne der Gesamtirrtumswahrscheinlichkeit statt, die Auswertungs- und Teststrategie wurde aber vollständig dargestellt.

Allerdings sind Hauptkomponentenanalyse und natürliche Splines eher unübliche Verfahren. Sie erfolgen datenabhängig, und die Interpretation oder grafische Darstellung erweist sich als schwierig.

Eine auch nur orientierende Ableitung von einzelstoffbezogenen Grenzwerten oder ähnlichem ist nicht möglich. Damit wird die Vermittlung der Hauptaussagen etwa an Entscheidungsträger erschwert, sodass man wieder vor der Gretchenfrage steht, wie viel Vereinfachung noch richtig ist.

## Literatur

Bhaskaran K, Gasparini A, Hajat S et al. (2013): Time series regression studies in environmental Epidemiology. In: *International Journal of Epidemiology* 42: 1187–1195.

Douwes J, Thorne P, Pearce N et al. (2003): Bioaerosol health effects and exposure assessment: progress and prospects. In: *Ann Occup Hyg* 47(3): 187–200.

Harrell FE (2001): *Regression modeling strategies*. New York.

Hartung J, Elpelt B, Klösener KH (1984): *Statistik. Lehr- u. Handbuch d. angewandten Statistik*. München, Wien.

Hoopmann M, Hehl O, Neisel F et al. (2006): Zusammenhang zwischen Bioaerosolen aus Tierhaltungsanlagen und asthmatischen Symptomen bei Kindern. In: *Gesundheitswesen* 68(8/09): 575–584

Hoopmann M, Csicsaky M, Schulze A et al. (2005): Gesundheitliche Bewertung von Bioaerosolen aus der Intensivtierhaltung in Niedersachsen. In: *UMID* 04: 3–6.

Iversen M (1999): Human health effects of dust exposure in animal confinement buildings. In: *Proceedings International symposium on dust control in animal production facilities*. Aarhus, Denmark. 30 May – 2 June 1999: 131–139.

Millner PD (2009): Bioaerosols associated with animal production operations. In: *Bioresource Technology* 100: 5379–85.

NLGA (2004): *AABEL: Atemwegserkrankungen und Allergien bei Einschulungskindern in einer ländlichen Region*. Endbericht. Hannover. <http://www.nlga.niedersachsen.de/download/77322/Endbericht.pdf> (Abrufdatum: 01.07.2014).

Nowak D (1998): Die Wirkung von Stallluftbestandteilen, insbesondere in Schweineställen, aus arbeitsmedizinischer Sicht. In: *Dtsch tierärztl Wschr* 105: 225–234.

O'Connor AM, Auvermann B, Bickett-Weddle et al. (2010): The Association between Proximity to Animal Feeding Operations and Community Health: A Systematic Review. *PLoS ONE* 5(3): e9530. doi:10.1371/journal.pone.0009530.

Radon K, Schulze A, Strien R et al. (2005a): Prevalence of respiratory symptoms and diseases in neighbours of large-scale farming in Northern Germany. In: *Pneumologie* 59 (12): 897–900.

Radon K, Schulze A, van Strien R (2005b): Atemwegsgesundheit und Allergiestatus bei jungen Erwachsenen in ländlichen Regionen Niedersachsens – Niedersächsische Lungenstudie. Abschlussbericht. München. <http://www.klinikum.uni-muenchen.de/Institut-und-Poliklinik-fuer-Arbeits-Sozial-und-Umweltmedizin/download/inhalt/Forschung/aumento/forschung/Abschlussbericht.pdf> (Abrufdatum: 01.07.2014).

Schlaud M, Salje A, Nischan P et al. (1998): MORBUS: Beobachtungspraxen in Niedersachsen. Bericht zur Erhebung in Süd-Oldenburg. In: Dtsch tierärztl Wschr 105: 235–240.

Sundberg, R (2002): Shrinkage regression. In: Encyclopedia of Environmetrics. Vol. 4. (Eds. El-Shaarawi AH, Piegorsch WW): 1994 – 98. Chichester. John Wiley & Sons, Ltd.

WHO (2004): Meta-Analysis of time-series studies of Particulate Matter (PM) and Ozone (O3). Report of a WHO task group. Kopenhagen. <http://apps.who.int/iris/handle/10665/107557> (Abrufdatum: 01.07.2014).

### **Kontakt**

Michael Hoopmann  
Niedersächsisches Landesgesundheitsamt  
Roesebeckstraße 4 – 6  
30449 Hannover  
E-Mail: Michael.Hoopmann[at]nnga.niedersachsen.de

[UBA]