Identifikation von Risikofaktoren für Stoffwechselerkrankungen durch Big Data-Analysen

Caspar Matzhold[1,2], Jana Lasser[1,2], Christa Egger-Danner[3], Franz Steininger[3], Birgit Fuerst-Waltl[4], Stefan Thurner[1,2,5,6], Peter Klimek[1,2]

[1]Complexity Science Hub Vienna, [2]Section for Science of Complex Systems, CeMSIIS, Medizinische Universität Wien, [3]ZuchtData EDV-Dienstleistungen GmbH, [4]Department für Nachhaltige Agrarsysteme, Institut für Nutztierwissenschaften, Universität für Bodenkultur Wien, [5]IIASA, [6]Santa Fe Institute, USA

Hintergrund und Zielsetzung

Die Digitalisierung in der Milchwirtschaft birgt das Versprechen Tierkrankheiten früher erkennen und gezielter verhindern zu können [1]. Dies wird insbesonders durch das Sammeln, Zusammenführen und Analysieren von Daten möglich, die häufig zu Routinezwecken ohnehin erfasst werden [2]. Das beinhaltet etwa Systeme welche Melkdaten, Fütterungsdaten oder andere Leistungsdaten aufzeichnen [3]. Die Identifikation von Risikofaktoren für Erkrankungen auf Basis solcher Daten wird jedoch erheblich dadurch erschwert, dass insbesonders häufige Stoffwechselerkrankungen oft durch ein Zusammenspiel von mehreren Faktoren gleichzeitig verursacht werden [4].

In dieser Arbeit präsentieren wir zum ersten Mal einen systematischen Zugang um die Effekte von Fütterungs- und anderen Betriebsmanagementpraktiken auf häufige Tiererkrankungen zu untersuchen. Wir betrachten dabei Ketose, Milchfieber, Lahmheit, Euterentzündungen, und Eierstockzysten. Das Ziel dabei ist es betriebsspezifische Risikoprofile für unterschiedliche Krankheiten erstellen zu können, d.h. an welcher Art von Betrieb welche Krankheiten häufiger oder weniger häufig als in anderen Betrieben vorkommen, und inwiefern einzelne Managementpraktiken in Zusammenhang mit diesem Risikoprofil stehen können.

Daten und Methoden

Daten. Wir verwenden Daten, welche in Vorprojekten gesammelt wurden. Dies beinhaltet insbesonders Daten aus dem „Efficient Cow" Projekt [5], aus dem umfangreiche Informationen zu Futteraufnahme, Milchleistung, Körpergewicht, Körpermaße, Tiergesundheit, sowie eine Vielzahl von Kennzahlen zum Betrieb vorliegen. Basierend auf Daten von 167 österreichischen Betrieben und 6.519 Kühen, können wir so den Zusammenhang von ca. 240 unterschiedlichen Variablen mit unterschiedlichen Krankheiten untersuchen.

In der Analyse gehen wir in zwei Schritten vor. Im ersten Schritt identifizieren wir Gruppen von ähnlichen Betrieben mittels einer Hauptkomponentenanalyse und untersuchen die Krankheitshäufigkeiten in diesen Gruppen. Im zweiten Schritt werden multivariate Regressionsmodelle ausgewertet um zu verstehen, welche Betriebsmerkmale am besten beobachtete Unterschiede in Krankheitshäufigkeiten erklären können. Im Folgenden werden diese beiden Auswertungsschritte kurz beschrieben.

Hauptkomponentenanalyse. Die Hauptkomponentenanalyse (HKA, oder auch bekannt als Principal Component Analysis, PCA) ist ein

mathematisches Verfahren der multivariaten Statistik zur Dimensionalitätsreduktion [6]. Sie identifiziert diejenigen Kombinationen von Variablen, welche die beobachtete Streuung in den Daten am besten erklären können. Diejenige Linearkombination von Variablen, welche die Varianz in den Daten am besten erklären kann, ist die sogenannte erste Hauptkomponente (PC1). Jede weitere Hauptkomponente steht orthogonal zu allen bisherigen Komponenten in einer Art und Weise, welche die übrig bleibende Varianz größtmöglich reduziert. Häufig wird neben der ersten auch die zweite Hauptkomponente (PC2) betrachtet, um eine zweidimensionale Beschreibung eines hochdimensionalen Datensatzes zu ermöglichen. Es ist zu erwarten, dass untereinander ähnliche Betriebe auch ähnliche Hauptkomponenten besitzen. Gruppen von ähnlichen Betrieben können somit durch Clusteringverfahren wie k-means im Hauptvektorenraum bestimmt werden: sie bilden „Haufen von Punkten“ mit ähnlichen Hauptkomponentenwerten. Innerhalb solcher Gruppen können wir dann Krankheitshäufigkeiten getrennt betrachten.

Multivariate Regression. Um herauszufinden welche Betriebsmerkmale die unterschiedlichen Krankheitshäufigkeiten am stärksten beeinflussen, betrachten wir multivariate Regressionsmodelle. Jede Erkrankung eines Tieres entspricht einer Beobachtung. Als Werte für die unabhängigen Variablen werden diejenigen Beobachtungen genommen, die zeitlich am nächsten vor der Erkrankung erfasst wurden. Als Kontrollgruppen wurden Tiere ohne Erkrankungen betrachtet mit den zuletzt beobachteten Werten. Für jede Variable und jede Krankheit wird dabei ein Modell ausge-

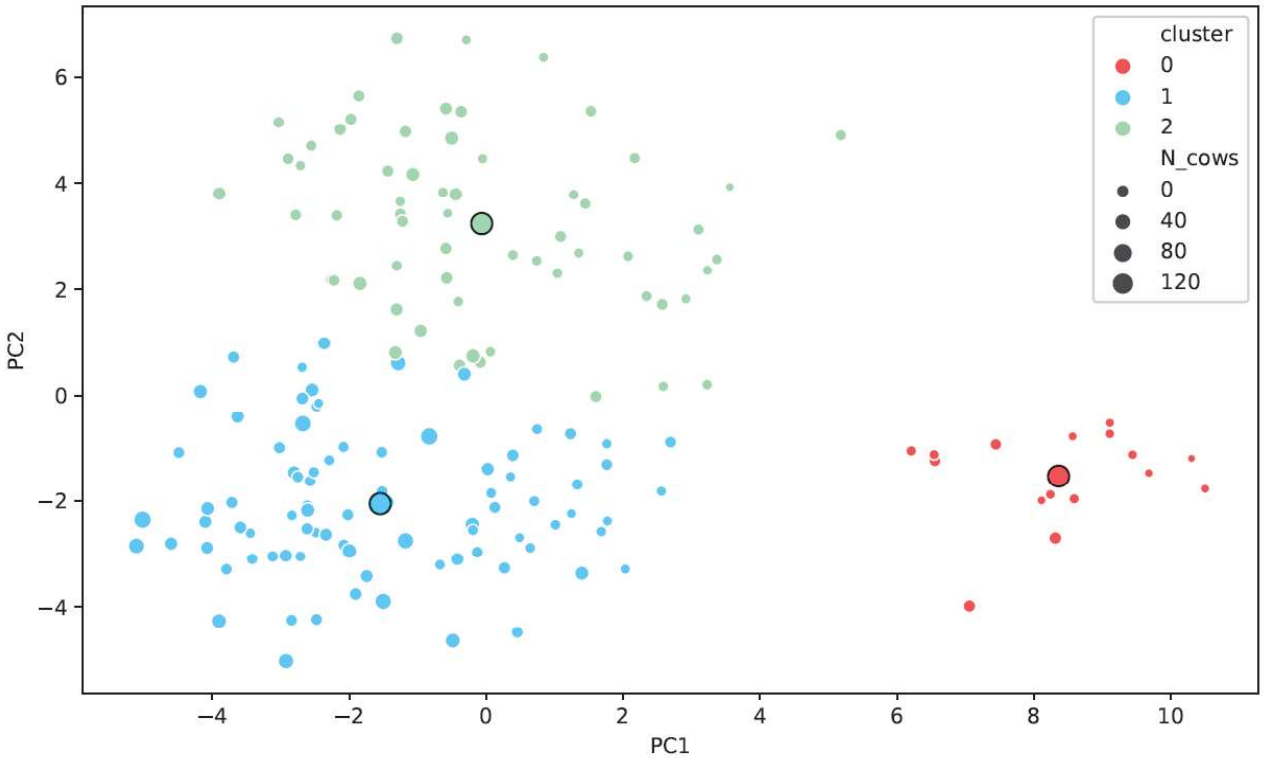

Abbildung 1: *Hauptkomponenten- und Clusteranalyse der Betriebe in den Efficient Cow Daten. Jeder Punkt entspricht einem Betrieb; die Größe der Punkte ist proportional zur Anzahl der Kühe an dem Betrieb. Es lassen sich drei Cluster von Betrieben erkennen, welche durch unterschiedliche Farben dargestellt sind.*

wertet, in dem für Laktation, Jahreszeit, Rasse, Region, sowie das Leistungsniveau des Betriebs adjustiert wird. Kategorische Variablen werden dabei „one-hot-codiert“, d.h. für jede Ausprägung der kategorischen Variable wird eine eigene Variable erstellt mit dem Wert von 1 falls die entsprechende Ausprägung bei der Beobachtung vorliegt, und 0 anderenfalls. Im Modell werden nur Variablen-Krankheitspaare betrachtet, für die zumindest 50 Beobachtungen vorlagen. Multiple Testkorrektur erfolgt nach Benjamini-Hochberg.

Resultate und Diskussion

Drei Betriebscluster. Die Ergebnisse der Hauptkomponentenanalyse der Efficient Cow Daten wird in Abbildung 1 gezeigt. Dabei betrachten wir die ersten beiden Hauptkomponenten (PC1 und PC2) und stellen sämtliche Betriebe in diesem Koordinatensystem dar. Jeder Betrieb entspricht dabei einem Punkt mit Größe proportional zur Anzahl seiner Kühe. Es ist klar offensichtlich, dass drei Cluster von Betrieben vorliegen (rot, blau, grün). Der rote Cluster hat dabei hohe Werte in PC1 während sich grün und blau in erster Linie in PC2 unterscheiden.

Charakterisierung der Cluster. Um zu verstehen, welche Betriebe in den drei Clustern liegen, kann man nun auswerten, wie stark unterschiedliche Variablen zu den beiden ersten Hauptkomponenten beitragen. Diese Ergeb-

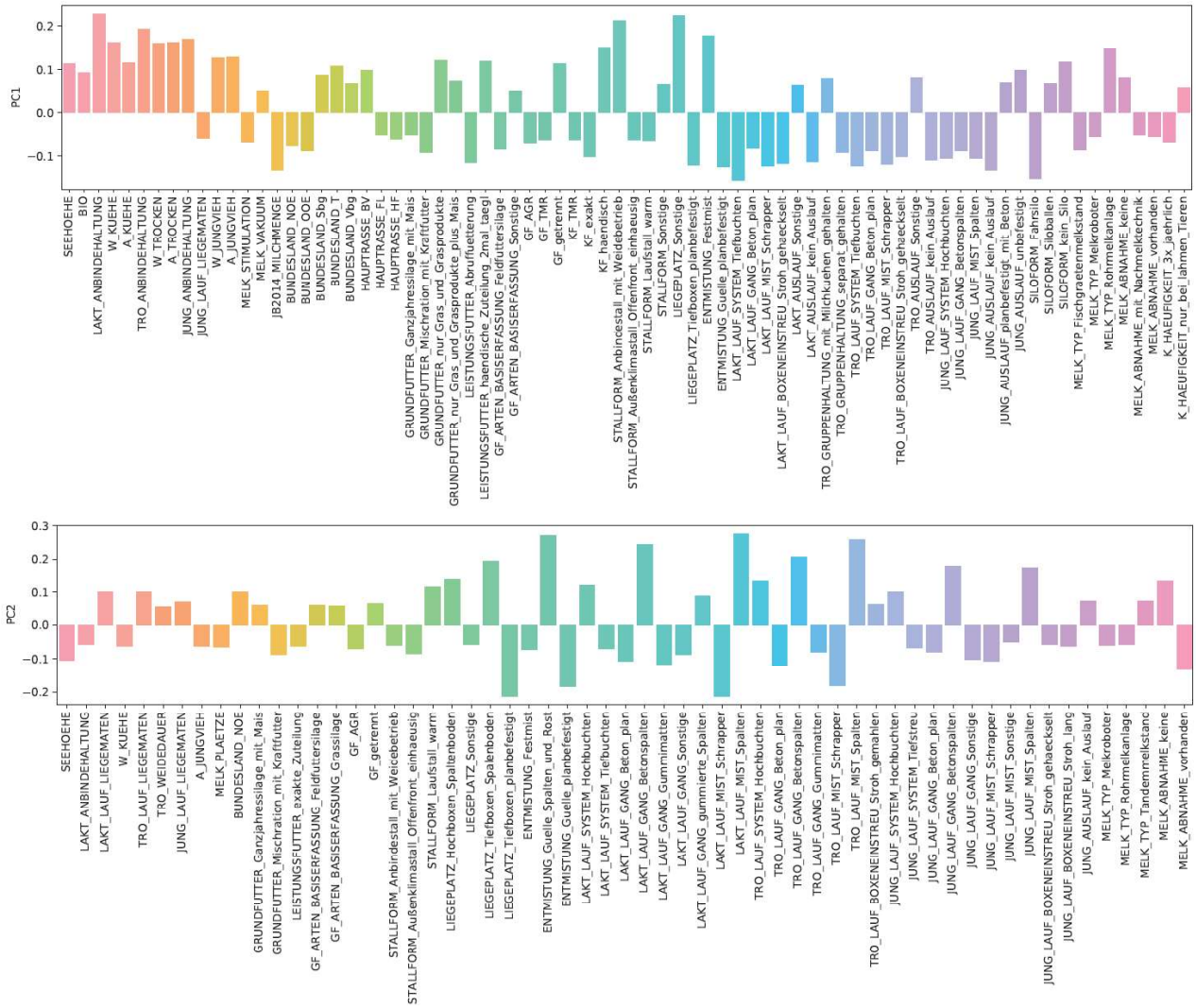

Abbildung 2: *Profile der ersten (oben) und zweiten Hauptkomponente (unten)*

nisse sind in Abbildung 2 dargestellt. Betriebe, bei denen PC1 hoch ist (roter Cluster) sind tendenziell kleinere Biobetriebe mit Anbindehaltung sowie Weidebetrieb und Alpungen, typischerweise auf größerer Seehöhe, mit Festmist, händischem Kraftfutter, Rohrmelkanlagen, etc. Im grünen und blauen Cluster sind damit verstärkt größere Betriebe, die sich anhand der Merkmale in PC2 unterscheiden. Der grüne Cluster hat im Vergleich zum Blauen hohe PC2 Werte, was insbesonders bei Betrieben mit Spaltenboden und Laufställen der Fall ist.

Krankheitsprofile der Cluster. Die drei Cluster unterschieden sich mitunter stark in der Häufigkeit einzelner Diagnosen. Insgesamt die niedrigsten Häufigkeiten findet man im roten Cluster. Die Krankheitshäufigkeiten in den anderen Clustern werden in Abbildung 3 gezeigt. Dabei zeigen sich in Cluster 1 erhöhte Fälle von Ketosen im Vergleich zu Cluster 2, während in Cluster 2 wiederum wesentlich mehr Fälle von Lahmheit im Vergleich zu Cluster 1 zu finden sind.

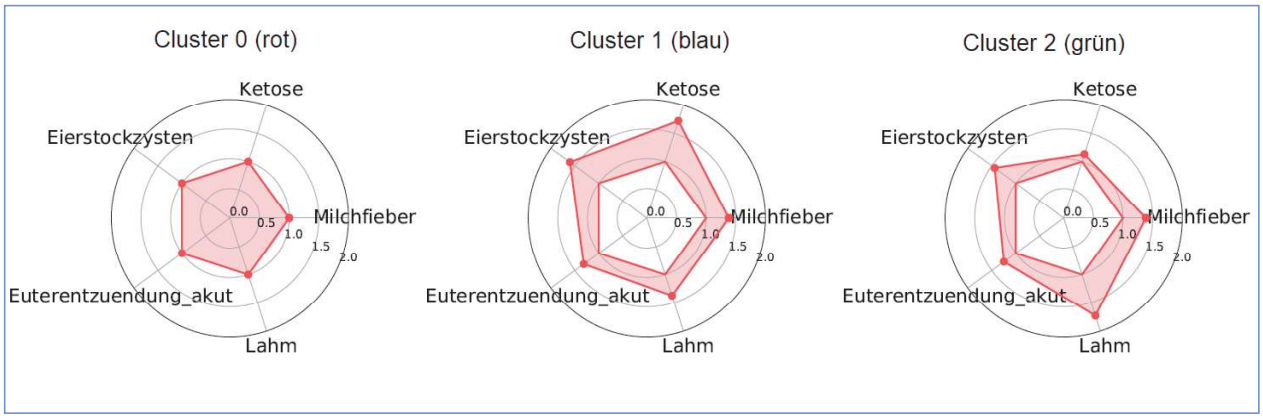

Abbildung 3: *Vergleich der Krankheitshäufigkeiten in den drei Clustern, mit Cluster 0 (rot) als Referenzgruppe. Für Cluster 1 und 2 werden jeweils die relativen Risiken relativ zu Cluster 0 angegeben.*

Risikofaktoren. Um diese Heterogenitäten besser zu verstehen, betrachten wir den Einfluss einzelner Risikofaktoren in multivariaten Regressionsmodellen. Die vollständige Auswertung dieser Daten ist momentan noch im Gange, folgende vorläufige Resultate treten dabei bis jetzt am stärksten heraus. Bezüglich Fütterung, finden wir zum Beispiel dass die Aufnahme von Grassilage prädiktiv für stark reduzierte Risiken von Eierstockzysten (Odds Ratio [OR] von 0.39, 95 % Konfidenzintervall [KI] von 0.31-0.50), Ketose (OR 0.42, KI 0.27-0.67) und Milchfieber (OR 0.69, KI 0.55-0.88) ist. Der Futtermittelanteil von Heu zeigte weiters protektive Effekte für Mastitis (OR 0.20, KI 0.085-0.46) aber auch für andere Krankheiten wie Lahmheit (OR 0.53, KI 0.40-0.71). Das Risiko für Lahmheit erhöht sich auch stark mit der Aufnahme von Kraftfutter (OR 15, KI 9.3-24) und Maissilage (OR 8.7, KI 6.7-11). Ketose zeigt sich stark reduziert in Betrieben mit Alpung von Jungtieren (OR 0.28, KI 0.18-0.43) sowie mit Boxeneinstreu von lang gehäckselten Stroh (OR 0.46, KI 0.33-0.64). Lahmheit wiederum nahm in Betrieben ohne Alpung zu (OR 4.6, KI 3.4-6.2) sowie in Betrieben mit Hochbuchten (OR 3.2, KI 2.8-3.7) während sich das Lahmheitsrisiko in Betrieben mit Tiefbuchten reduzierte (OR 0.44, KI 0.40-0.50).

Das Ergebnis eines stark erhöhten Lahmheitsrisikos in Betrieben mit Hochbuchten ist damit konsistent mit dem Ergebnis einer Häufung von Lahmheiten in Betriebscluster 2 (grün), wo tendenziell eher Hochbuchten ein-

gesetzt werden. Dies führt exemplarisch vor, wie unsere Ergebnisse verwendet werden können um Risikoprofile für einzelne Betriebe erstellen zu können, und die wahrscheinlichsten Treiber von Unterschieden in den Krankheitshäufigkeiten identifizieren zu können.

Wir bitten zu beachten, dass dieses Dokument work in progress beschreibt und kleinere Verfeinerungen in der Auswertungs- und Berechnungsmethodik noch zu erwarten sind, bevor endgültige Resultate in einer wissenschaftlichen Publikation angestrebt werden.

Schlussbemerkungen

Bauernhöfe des 21. Jahrhunderts werden immer mehr zu digital integrierten Sensorsystemen. Mit unserem Zugang wollen wir zeigen, wie man die Informationen solcher Systeme zur Erhöhung des Tierwohls verwenden kann indem man diejenigen Managementpraktiken identifiziert, deren Anwendung zu den größten Risikoreduktionen einer Vielzahl von Krankheiten führt.

Danksagung

Diese Arbeit wurde im Rahmen des COMET-Projekts D4Dairy (Digitalisation, Data integration, Detection and Decision support in Dairying, Projektnummer: 872039) durchgeführt, das vom BMK, BMDW und den Ländern Niederösterreich und Wien im Rahmen von COMET-Kompetenzzentren für exzellente Technologien unterstützt wird. Das COMET-Programm wird von der FFG abgewickelt.

Für diese Studie wurden Ergebnisse aus dem Projekt "Efficient Cow" verwendet, das vom österreichischen Bundesministerium für Land- und Forstwirtschaft, Umwelt und Wasserwirtschaft, den Bundesländern, der ZAR mit Mitgliedsorganisationen und weiteren Partnerorganisationen unterstützt wurde.

Literatur

[1] C. Lokhorst, R. de Mol und C. Kamphuis, „Invited review: Big Data in precision dairy farming," Animal, Bd. 13, Nr. 7, pp. 1519-1528, 2019.

[2] T. Norton und D. Berckmans, „Developing precision livestock farming tools for precision dairy farming," Animal Frontiers, Bd. 7, Nr. 1, pp. 18-23, 2017.

[3] J. Bewley, „Precision dairy farming: advanced analysis solutions for future profitability.," in Proceedings of the first North American conference on precision dairy management, Toronto, Canada, 2010.

[4] A. Stone, „Precision dairy farming technology solutions for detecting dairy cow disease to improve dairy cow well-being," Theses and Dissertations--Animal and Food, Bd. 70, 2017.

[5] F. Steininger, C. Fuerst, B. Fuerst-Waltl, L. Gruber, M. Mayerhofer, M. Ledinek, R. Weissensteiner, W. Zollitsch, K. Zottl und C. Egger-Danner, „Efficient cow: Strategies for on-farm collecting of Efficient cow: Strategies for on-farm collecting of phenotypes for efficiency traits," ICAR Technical Series, Bd. 19, pp. 167-174, 2015.

[6] T. Hastie, R. Tibshirani, Friedman und Jerome, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Series in Statistics, 2001.