Analyse eines zweistufigen, regionalen Clusteralgorithmus am Beispiel der Verbundenen Wohngebäudeversicherung

Zusammenfassung der Diplomarbeit an der Hochschule Zittau/Görlitz

Maria Kiseleva

Motivation und Ziel

Die Regionalisierung ist in der Versicherungswirtschaft ein wichtiges Thema. Regionale Tarife in der Kfz-Versicherung, der Gebäude- und Hausratversicherung sind seit langem bekannt. Schätzungen von Naturkatastrophen, wie z.B. Sturmschäden, sind ebenfalls regional abhängig.

Das Ziel dieser Diplomarbeit ist die Entwicklung und Untersuchung eines neuen Algorithmus für die regionale Analyse einer vorgegebenen Zielgröße. Er dient der Bildung von regionalen Clustern, die sich in Abhängigkeit der Zielgröße unterscheiden. Das Verfahren basiert auf numerischen und statistischen Methoden und stellt eine neue Vorgehensweise der Regionalisierung vor.

Regionale Analyse

Für regionale Analysen wird die Annahme getroffen, dass sich die jeweilige Schadenzielgröße regional differenzieren lässt, d.h. dass die Region einen Einfluss auf den Schadenverlauf der Risiken hat. Diese Arbeit betrachtet einen Algorithmus, der das Ziel verfolgt, Regionen in Zonen mit unterschiedlichem Schadenverlauf einzuteilen.

Als Datenbasis für den Algorithmus dienen die Schadendaten auf regionaler Ebene, z. B. Schadendaten je Postleitzahlgebiet (PLZ). Gemäß dem Gesetz der großen Zahlen, werden die PLZ zunächst in geographisch zusammenhängende Gebiete gegliedert, sodass jedes der Gebiete eine

Mindestbestandsgröße (MBG) an Risiken aufweist. Darauf aufbauend werden die Gebiete zu einer bestimmten Anzahl an Clustern zusammengefasst. Die Vorgehensweise kann in vier Schritten dargestellt werden:

1. *Aggregation der Daten*: Die Daten werden zunächst auf PLZ-Ebene aggregiert. Je PLZ p_i müssen die Geokoordinaten (Längen- und Breitengrad), die Anzahl der Risiken n_i, die Gewichtung für die Schadenzielgröße g_i und die Schadenzielgröße Z_i ermittelt werden.
2. *Bildung statistisch repräsentativer Zonen*: Für die Verdichtung der PLZ muss eine MBG angegeben werden. Ausgehend von der PLZ mit der minimalen Anzahl an Risiken werden die benachbarten PLZ solange zusammengefasst bis die MBG erreicht wird. Für die weiteren Schritten wird die Schadenzielgröße in den PLZ-Clustern $\tilde{Z}_i$ als gewichteter Mittelwert der Schadenzielgrößen der zugeordneten PLZ berechnet.
3. *Glättung der Schadenzielgröße*: Um große Schwankungen der Schadenzielgröße zwischen benachbarten PLZ-Clustern zu reduzieren, wird die Schadenzielgröße über die Cluster geglättet. Als Kriterium für die Glättung dient eine zweite MBG, die größer als die erste MBG sein soll. Die Glättung erfolgt nach einem Credibility-Ansatz:

   $$\tilde{Z}_i^G = \lambda \tilde{Z}_i + (1 - \lambda)\tilde{Z}_j ,$$

   wobei $\tilde{Z}_j$ – die Schadenzielgröße im Nachbar-Cluster des PLZ-Cluster *i* ist; $\lambda = \sqrt{\frac{\tilde{g}_i}{\tilde{g}_i + \tilde{g}_j}}$ – der Credibilityfaktor bezogen auf die Gewichtung $\tilde{g}_i$ der Schadenzielgröße in PLZ-Cluster *i*.
4. *Clusterung der neuentstandenen Gebiete*: Jeder einzelnen PLZ wird die geglättete Schadenzielgröße aus dem entsprechenden PLZ-Cluster zugeordnet. Die PLZ werden nach dem Wert der Schadenzielgröße mit Hilfe eines Cluster-Verfahrens in die gewünschte Anzahl an Zonen eingeteilt. Es wurde nicht festgelegt, welcher Cluster-Algorithmus zu verwenden ist. In dieser Arbeit werden die Zonen mit Hilfe des Exakten Cluster-Algorithmus gebildet.

Während der erste und der letzte Schritt des Algorithmus für die statistische Clusterbildung üblich sind, müssen für den zweiten und dritten Schritt zunächst die Fragen hinsichtlich einer geeigneten MBG bzw. Nachbarschaftsdefinition beantwortet werden.

Die Mindestbestandsgröße m kann nach dem Zentralen Grenzwertsatz bestimmt werden1:

$$m \geq \frac{u_{1-\frac{\alpha}{2}}^2}{\varepsilon^2}\frac{\sigma^2}{\mu^2},$$

wobei $u_{1-\frac{\alpha}{2}}$ - das zweiseitige α-Quantil der Standardnormalverteilung ist; ε - die akzeptierte Abweichung zum Mittelwert der Zielgröße μ; σ^2 - die Varianz der Zielgröße.

Da zwei MBG für den Algorithmus notwendig sind, werden für die zweite MBG die Parameter α und ε strenger gesetzt. Dadurch wird die MBG für den dritten Schritt größer als die erste MBG.

Für die Bestimmung der Nachbarschaftsrelationen werden die Geokoordinaten der einzelnen PLZ benötigt. Sie bieten die Möglichkeit die PLZ als Punkte auf einer Ebene zu betrachten. Eine Menge von Punkten kann durch einen Graph ersetzt werden. Wenn dieser Graph ein vollständiger Graph ist, d.h. jede PLZ ist eine Nachbar-PLZ zu allen anderen PLZ, wird der Algorithmus beim zweiten Schritt die PLZ-Cluster nach der minimalen Entfernung zwischen den PLZ bilden.

In der Realität sind die PLZ jedoch nur durch eine gemeinsame Grenze verbunden. Deswegen ist die Anzahl an Nachbar-PLZ begrenzt. Eine zweite Variante der Nachbarschaftsrelation berücksichtigt diesen Fakt und ist deswegen für die praktische Anwendung zu empfehlen.

---

[1] I. B. Hossack, J. H. Polland and B. Zehnwirth, Introductory statistics with applications in general insurance, Cambridge University Press; Auflage: 2, 1999.

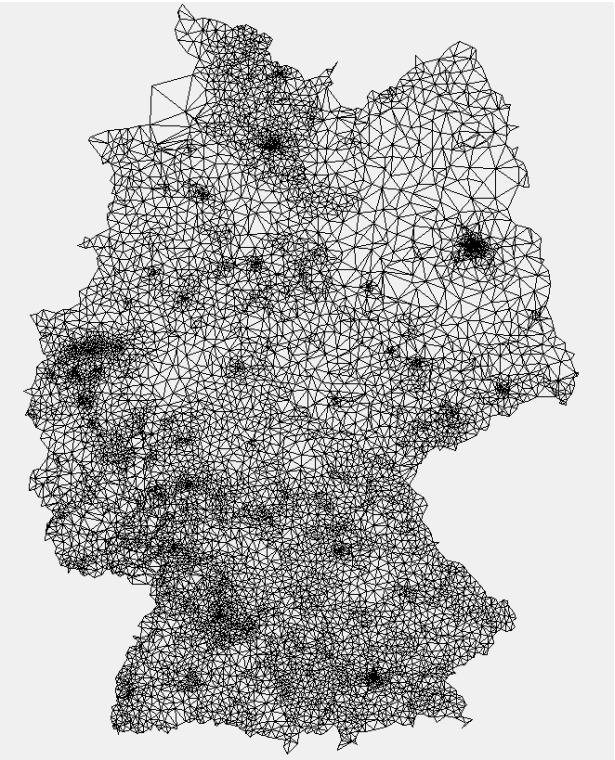

Abb. 1 Mathematische Darstellung der PLZ in Deutschland mit Hilfe des Dreiecksnetzes

Bei dieser zweiten Variante werden die PLZ durch ein Dreiecksnetz verbunden. Dieses wird mit Hilfe der Delaunay-Triangulation[2] gebildet. In diesem Fall definiert der Algorithmus diejenigen PLZ als Nachbar-PLZ, die durch eine gemeinsame Kante verbunden sind. Falls zwei PLZ durch einen Weg der Länge *n* verbunden sind, spricht man von einer Nachbarschaft der Potenz *n*.

Diese Variante erzielt bei der Zusammenfassung, dass die MBG erreicht wird und gleichzeitig die Varianz der Schadenzielgröße innerhalb der PLZ-Gruppe minimal ist. So entsteht eine zusätzliche Brücke zur Clusteranalyse, da dort die Zusammenfassung der PLZ in Zonen ebenfalls hinsichtlich der Minimierung der Varianz innerhalb der Zonen erfolgt.

Modellierung

Der Algorithmus wurde anhand von Daten von Leitungswasserschäden in der Verbundenen Wohngebäudeversicherung (VGV), die ab Statistikjahr 2005 vorliegen, getestet. Die Daten wurden in fünf Zeitperioden eingeteilt: 2005-2008, 2006-2009, 2007-2010, 2008-2011, 2009-2012. Das bietet

[2] Sieh F. P. Preparata and M. I. Shamos, Computational Geometry: An Introduction, New York - Berlin - Heidelberg - Tokyo: Springer, 1985; R. Klein, Algorithmische Geometrie: Grundlagen, Methoden, Anwendungen, Berlin Heidelberg: Springer, 2005.

sowohl die Möglichkeit die Modelle an einzelnen Statistikjahren 2009, 2010, …, 2013 zu validieren, als auch die zeitliche Stabilität zu überprüfen.

Für jede der oben genannten Zeitperioden wurden mit Hilfe des Algorithmus drei Zonierungsmodelle, bei denen die PLZ in drei, vier oder fünf Zonen eingeteilt wurden, berechnet und an den jeweils nicht in die Modellbildung eingeflossenen Statistikjahren validiert. Die Differenzierung der Schadenkennzahlen durch die gebildeten Zonen konnte bei allen Modellen bestätigt werden. Zudem konnte die zeitliche Stabilität der Zonen nachgewissen werden. Der Vergleich des vier Zonen-Modells mit der GDV - Einteilung zeigte, dass das neu ermittelte Modell einen deutlich größeren Spreiz des Schadensatzes zwischen den Zonen aufweist.

Als zusätzliche Validierung wurde das vier Zonen-Modell schließlich als Merkmal in einem multivariaten Risikomodell verwendet. Die Modellierung erfolgte mit Hilfe eines Verallgemeinerten Linearen Modells. Als Zielgröße diente der Schadensatz (= Schadenaufwand/Versicherungssumme). Die resultierenden Risikofaktoren bestätigten ebenfalls die starke Differenzierung des Leitungswasser-Risikos durch die gebildeten regionalen Zonen.

Zusammenfassung

Der im Rahmen dieser Diplomarbeit entwickelte und vorgestellte Zonierungsalgorithmus verknüpft verschiedene statistische und numerische Methoden. Ein Vorteil dieses Algorithmus ist, dass er für beliebige Zielgrößen und in unterschiedlichen Bereichen anwendbar ist. Die praktische Anwendung des Algorithmus auf empirische Daten, zeigt die Stabilität und Validität der Ergebnisse. Er kann somit unmittelbar in der Praxis eingesetzt werden.