# Nominalskaliert interpretierte Befragungen in der Verdachtsgenerierung

Pitlik L, 2012-01-22

## Einführung

Verdachtsgenerierung kann anhand von mehreren Interpretationsebenen durchgeführt werden (vgl. outlier detection: <http://miau.gau.hu/miau/162/outliers.docx>). Zum ersten kann man zielgerichtete Bewertungen durchführen, falls man auf jede Frage eine priorisierte Antwort feststellen kann, bzw. sich die Akzeptierbarkeit der Antwortmöglichkeiten pro Frage als Rangfolge verstehen lässt (vgl. <http://miau.gau.hu/miau/156/bericht_frageboegen.docx>). Zum zweiten kann man auf monotone Zusammenhänge der Antwortoptionen scheinbar verzichten (vgl. <http://miau.gau.hu/miau/159/93_2.docx>), um dafür aber spezielle Modellkontrollmöglichkeiten zu gewinnen. Zum dritten können alle Fragebogeninformation als nominalskaliert verstanden, damit man nach einer Art kombinatorische Seltenheit als Verdacht suchen kann (vgl. <http://miau.gau.hu/miau/162/suspicion_banking2.docx>). In diesem Beitrag wird das letztere kurz beschrieben, da die davor sind bereits dokumentiert worden.

## Verdachtsgenerierung als nominalskaliertes Problem

Nominalskaliert lassen sich alle Informationsarten und –Inhalte verstehen. Nominalskaliert sind ab ovo Geschlechtsdaten, Berufsformen, Standorte etc. Bei anderen Variablen (Fragen) darf man nicht einmal aus monotonen Zusammenhängen zwischen den erfassten Antworten und jegliche Art von Verdachtsmomenten (vgl. Optimum-Logik oder Oszillation). Falls man jedes Datum als eine selbständige Identität akzeptiert, lassen sich in einer Datenbasis alle alphanumerische Charakterketten mit einer fortlaufenden positiven, Ganzzahl (1,…, n) als Code ersetzt werden. Alle weiteren Schritte werden sich auf diese Codes beziehen.

## Modellierungslogik

In der Ähnlichkeitsanalyse braucht man (quasi) Rangzahlen als Inputwerte. Die oben eingeführten Codes besitzen die notwendigen Eigenschaften. Ähnlichkeitsanalysen basieren auf LP-Modelle, in denen man den Treppenfunktionen gegenüber Restriktionen formulieren dar, aber man muss es nicht unbedingt tun. Die COCO-Modelle, bei denen keine Restriktionen bezüglich der Treppeneigenschaften vorliegen werden als COCO-MCM Modelle genannt, da man die Treppenstufenparameter quasi zufällig festlegt (vgl. Monte-Carlo Verfahren). Eine COCO-MCM-Konstrukt ist in der Lage beliebige kombinatorische Situationen abzubilden, da man nicht weiteres sucht als die Zahlenkombinationen, deren Summe im Endergebnis den Wert der abhängigen Variable pro Objekt am genauesten reflektieren kann. In einer Verdachtsgenerierung werden diverse Verdachtssichten untersucht: jede Inputvariable wird anhand von allen übriggebliebenen Variablen modelliert. Das heißt: im Falle einer Objekt-Attribut-Matrix mit 93 Personen als Befragten und 40 Fragen braucht man in diesem Fall 40 selbständige Modelle auf die Beine zu stellen. Die Abweichungen der 40 Norm- oder Schätzwerte und der passenden IST-Werte werden dann mit sich multipliziert (damit die Vorzeichen verschwinden) um eine Art Verdachtsindex bilden zu können. Die gleiche Aggregationslogik gilt auch für die Fragen. Das Verfahren liefert letzten Endes nichts anderes als zwei Vektoren von Verdachtsindizes: einmal für die Befragten und einmal für die Fragen selbst.

## Ergebnisse

Tabelle 1 zeigt Verdachtssichten noch mit Verdachtsvorzeichen. Rechts und unten sind die aggregierte Verdachtsindexwerte zu sehen. Die Zeilen sind die Befragten, die Spalten sind die Fragen.

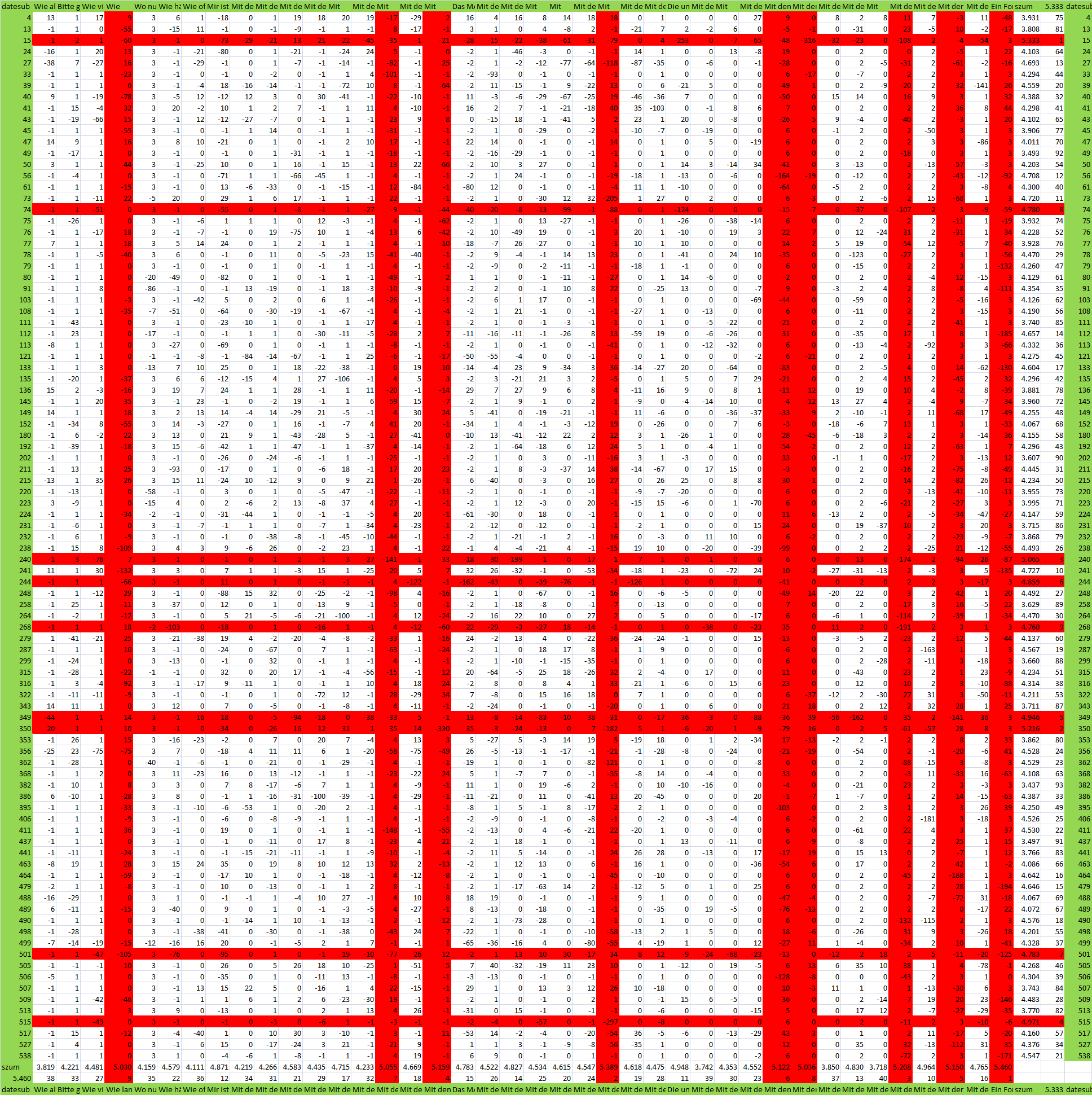


Tabelle 1: Verdachtssichten (Quelle: eigene Abbildungen)

Die Abbildung 1 und 2 zeigen die Verteilung der logarithmisch transformierte Verdachtsmomente. Vergleicht man die Spitzenverdachtsmomente mit den Ergebnissen anderer Verfahren (s. Einführung), dann lässt sich feststellen, dass die am meisten unlogisch, inkonsistent behandelte Frage die letzte Frage ist. Dies verstärkt im Weiteren die Interpretation, dass die letzten Fragen in einem Fragebogen quasi als Störfaktor angesehen werden, und auf diese Frage kann man kaum vernünftige Antworten erwarten. Außerdem vertreten sich die Fragen, die am Anfang eines Fragebogens zu finden sind, „signifikant“ häufiger in der Gruppe mit Spitzenverdachtsmomenten als am anderen Ende der Verteilung.

Abbildung 1: Verteilung der Verdachtsmomente bei den untersuchten Objekten (Qulle: eigene Berechnungen)

Abbildung 2: Verteilung der Verdachtsmomente bei den untersuchten Variablen (Qulle: eigene Berechnungen)

Bezüglich der Stabilität der objektbezogene Verdachtsmomente sind ähnliche Aussagen zu treffen. 8 von den ersten 10 Objekten sind in den anderen Verfahren ebenso verdächtig. Verdacht entsteht in einer S-förmigen Verteilung, welche mehrere charakteristische Gruppierungsschwellenwerte besitzt (vgl. frühere Studien).

## Zusammenfassung

Über Verdacht (inkonsistentes Verhalten) lässt sich also fundiert sprechen. Abhängig davon, was man mit den Rohinformationen erzielen wollte, sind ganz konkrete Aktionen anhand der Verdachtsmomente zu starten. Da die Berechnungen automatisch ablaufen und gut skalierbar sind, können Verdachtsbehandlungen auch real-time stattfinden. Ähnliche Analysen mit gleich positiven Erfahrungen wurden bereits im Banksektor (vgl. Bankkartentransaktionen) und im Bereich des finanziellen Bewusstseins durchgeführt (<http://miau.gau.hu/miau2009/index.php3?x=e37>)…