Anwendung der Methode der mehrfachen Regression für die Analyse von Belastungskurven [Fortsetzung]

Autor(en): **Védère, Elie**

Objekttyp: **Article**

Zeitschrift: **Bulletin des Schweizerischen Elektrotechnischen Vereins : gemeinsames Publikationsorgan des Schweizerischen Elektrotechnischen Vereins (SEV) und des Verbandes Schweizerischer Elektrizitätswerke (VSE)**

Band (Jahr): **53 (1962)**

Heft 20

PDF erstellt am: **21.09.2024**

Persistenter Link: https://doi.org/10.5169/seals-916980

Anwendung der Methode der mehrfachen Regression für die Analyse von Belastungskurven

Von *Elie Védère*, Paris

(Fortsetzung aus Nr. 19, S. 924)

2. Gültigkeit der Resultate bei der Verwendung der mehrfachen Regression zur Analyse von Belastungskurven

2.0. Über die Gültigkeit der Resultate einer Untersuchung mit Hilfe der mehrfachen Regression (wie überhaupt von jeder Analyse) lässt sich folgendes sagen:

1. Sie hängt von den Annahmen ab, die getroffen werden mussten (logisch sollten diese Annahmen derart sein, dass die Fehler am kleinsten werden).
2. Sie wird durch Messungen oder Berechnungen bestimmt.

Diese Punkte sollen nun nacheinander näher erörtert werden.

2.1. A — Annahmen bei der Verwendung der mehrfachen Regression zur Analyse von Belastungskurven

2.10. Diese Annahmen sind im wesentlichen die folgenden:

a) Art der vorzunehmenden Zerlegung
b) Wahl der unabhängigen Variablen
c) Wahl der «physikalischen» Grundlagen der N Gleichungen
d) Wahl der Anzahl Gleichungen (N)
e) Entschluss betr. Einführung oder Nichteinführung eines konstanten Gliedes in die Regressionsgleichungen.

2.11. a) *Art der vorzunehmenden Zerlegung*

Die Zerlegung der Belastung oder des Verbrauches soll zu Belastungs- oder Verbrauchswerten führen, die von gleicher Grössenordnung sind. Eine zu kleine Teilbelastung oder ein zu kleiner Teilverbrauch würde in der natürlichen Streuung der totalen Belastung oder des totalen Verbrauches «untergehen».

Es gibt hier eine fundamentale Regel, deren Missachtung zu Ergebnissen führen würde, die mit allzugrossen Fehlern behaftet wären;

— Beispiel: bei der Zerlegung einer Belastung in ihre Komponenten (Grundbelastung und von meteorologischen Faktoren abhängige Schwankungen), sollen die Faktoren «Windgeschwindigkeit» oder «Luftfeuchtigkeit» nur dann berücksichtigt werden, wenn sie namhafte Belastungsschwankungen zur Folge haben.

— Weiteres Beispiel: bei einer Analyse des Verbrauches mit dem Ziel, dem mittleren Verbrauch der einzelnen Verbrauchsapparate zu ermitteln, wird es meistens schwer fallen, neben dem Verbrauch grösserer Apparate, wie z. B. der Kochherde, auch denjenigen einzelnen Kleinapparate (wie z. B. der Radioapparate), zu erfassen.

Es steht jedoch fest, dass unter sonst gleichen Bedingungen, die Präzision der Resultate mit der Zahl der Gleichungen steigt.

2.12. b) *Wahl der unabhängigen Variablen*

Mit der Festlegung der Art der Zerlegung (2.11) ist noch keineswegs die Wahl der unabhängigen Variablen bestimmt. In den meisten Fällen ist die Auswahl reichlich, was aus folgenden Beispielen hervorgeht:

Bei Untersuchungen in der Zeit können die meteorologischen Faktoren durch verschiedenen Variablen dargestellt werden. Um die Linearität der Regression zu wahren, haben wir bereits darauf hingewiesen, dass statt der Tageshelle deren Logarithmus zu wählen sei. Welcher Wert soll aber eingesetzt werden? Ist es die mittlere Beleuchtungsstärke im ganzen Netz? Gewiss, aber wie soll sie dann gewichtet werden? Und die Temperatur, ist es diejenige des betreffenden Tages, oder sollte man nicht auch die Temperatur des Vortages (oder sogar der vorangehenden Tage) mitberücksichtigen?

Was die Untersuchungen nach geographischen Zonen betrifft, hat man u. a. für den Anteil der Haushaltungen die Wahl zwischen «installierte Leistung», «Anzahl Abonnenten» und andere mehr. Die Wahl der Variablen ist oft eine intuitive Angelegenheit; sie kann aber auch gelenkt werden [9]):

2.121. Hat man verschiedene Gruppen von Variablen gewählt, so genügt es (nachträglich), die Streuung von Y zu untersuchen und diejenige Gruppe zu behalten, die zur kleinsten Reststreuung führt.

Es sei hier daran erinnert, dass für jede Beobachtung p unter den N Beobachtungen eine Reststreuung $\varepsilon_p - Y_p = Y'_p$ [s. 0.24 Gleichung (3)] definiert werden kann, wobei die Y'_p die N «geschätzten» Werte sind.

Die Reststreuung ist durch die Formel gegeben:

$$V = \frac{1}{N - n - 1} \cdot \sum_{p=1}^{p=N} \varepsilon_p{}^2 .$$

2.122. Soll eine Analyse mit Hilfe der mehrfachen Regression zu gutem Ziel führen, so darf

[9]) Die verwendeten Variablen (s. die Beispiele in Beilage I) zeugen von ihrer oft willkürlichen Auswahl, was in jedem einzelnen Falle eine Überprüfung der Gültigkeit der getroffenen Annahmen bedingt.

zwischen den unabhängigen Variablen keine nennenswerte Korrelation bestehen.

Ziel der Analyse ist ja eine Zerlegung der Belastung, d. h. ihre Unterteilung in eine Anzahl voneinander eindeutig verschiedene Teillasten. Die Untersuchung einer mehrfachen Regression, bei der die «unabhängigen» Variablen X nicht alle untereinander streng unabhängig wären, dürfte mathematisch nicht unmöglich sein; sie würde aber nicht mit Bestimmtheit zur Zerlegung führen, die wir eben suchen.

Es muss also vorerst geprüft werden, ob die «unabhängigen» Variablen wirklich von einander unabhängig sind, was zu den klassischen Aufgaben der mathematischen Statistik gehört.

2.123. In bezug auf ihre Unabhängigkeit müssen diese Variablen folgenden Bedingungen genügen:

- Sie dürfen durch keine lineare Beziehung der Form $X_1 = KX_2 + a$ gebunden sein, und auch nicht einen Zusammenhang der Art $X_1 + X_2 + X_3 + ... = 100\%$ aufweisen.
- Sie sollen keine enge Korrelation aufweisen, d. h. wenn b_1 genau bestimmt werden soll, darf der mehrfache Korrelationskoeffizient zwischen X_1 und den anderen unabhängigen Variablen nur sehr klein sein (diese Bedingung kann aber fallen gelassen werden, wenn b_1 nur zur Bestimmung von Y dient);
- die zu jedem Paar unabhängiger Variablen gehörenden einfachen Korrelationskoeffizienten können berechnet werden: ein hoher Korrelationskoeffizient (z. B. höher als 0,5) ist ein Zeichen dafür, dass eine genaue Unterscheidung der Wirkung der zwei betrachteten unabhängigen Variablen nicht möglich sein wird.
- Schliesslich können noch die diagonalen Elemente C_{ii} der inversen Matrize (0.26 und 2.211) einen Hinweis für die Abhängigkeit liefern, indem das grösste dieser Elemente dem Repressionskoeffizient mit der grössten Streuung entspricht (s. unter 2.212 - c).

2.124. Das Vorhandensein einer hohen Korrelation zwischen unabhängigen Variablen wird dazu veranlassen, andere Variablen zu suchen oder zu untersuchen, ob nicht gerade die physikalischen Grundlagen der Gleichungen Ursache dieser Korrelation sind (s. unter 2.131).

Besteht trotz allem noch eine Korrelation zwischen zwei Variablen, so heisst das nicht, dass diese Korrelation die anderen Resultate der Analyse beeinflussen könnte. Nur die Unterscheidung zwischen den Wirkungen der zwei in Frage stehenden Variablen ist unsicher. Es bleibt dann nichts anderes übrig, als die Elemente dieser zwei Variablen zusammenzufassen und auf die a priori gewünschte Unterscheidung zu verzichten. Die Untersuchung wird sich aber trotzdem gelohnt haben, weil sie die Nutzlosigkeit einer oder mehrerer Variablen aufgezeigt hat. (Anmerkung der Redaktion: Ein solcher Fall könnte beispielsweise bei einer Untersuchung des Haushaltverbrauches auftreten, in einem Netze wo die Heisswasserspeicher nicht gesperrt sind, so dass zwischen deren Verbrauch [bzw. Belastung] und demjenigen der Kochherde sicher eine merkliche Korrelation bestehen dürfte.)

2.13. c) *Wahl der physikalischen Grundlage der N Gleichungen*

Hier wird eines der heikelsten Probleme angeschnitten, die es bei der Anwendung der Methode gibt, wenn diese zu einer Analyse nach geographischen Zonen führt. Wenn dieser Wahl bei einer Analyse in der Zeit keine allzugrosse Bedeutung zukommt, weil jede Gleichung einem bestimmten Tage entsprechen wird (immerhin ist dabei zu beachten, dass Feiertage und Werktage nicht gleiche Charakteristiken aufweisen), so lässt dagegen die Aufstellung der Gleichungen nach geographischen Zonen viel mehr Möglichkeiten zu.

2.131. Vorerst kann eine Abhängigkeit zwischen zwei Variablen (s. unter 2.122) zu einer Umgruppierung der geographischen Zonen führen, nach welcher dann die unerwünschten Abhängigkeiten mit etwas Glück verschwinden.

Bei der Analyse der Belastungskurve eines Netzes, kann zum Beispiel die Industrie-Belastung in den verschiedenen Netzteilen ohne weiteres als unabhängig von der Belastung der Haushaltungen und des Gewerbes betrachtet werden. Diese zwei Teilbelastungen sind aber in grossen Verteilzonen in einem gewissen Masse voneinander abhängig. Um die Unabhängigkeit der Variablen zu verbessern, ist es also nötig, kleinere geographische Zonen mit verschiedenen Merkmalen vorzusehen (wie Geschäftszentren von grossen Städten und Wohnzonen).

2.132. Anderseits sollten die Gruppen, unter die die Gesamtbelastung aufzuteilen ist, relativ *homogen* sein.

Es ist klar, dass wenn die Abonnenten oder Abonnentengruppen der graphischen Zone p Belastungen, bzw. Verbrauchszahlen aufwiesen, die den Werten $X_{1p}...X_{np}$ streng proportional wären, $n + 1$ Gleichungen genügen würden, um die $n + 1$ Wurzeln a, $b_1...b_n$ eindeutig zu bestimmen und, dass in jeder Gleichung (3) ε_p gleich Null wäre.

Dies würde bedeuten, dass alle Abonnenten oder Abonnentengruppen unter gleichen Bedingungen (d.h. für gleiche Werte der unabhängigen Variablen) sich identisch verhielten.

Je näher man diesem idealen Bild kommt, desto besser werden die Resultate; darin liegt das Kriterium der Homogenität.

Man wird sich also bemühen, Abonnenten oder Abonnentengruppen zu wählen, die in bezug auf den Gegenstand der Untersuchung (z. B. «Haushaltungen» oder «Kochherde») ähnliche Reaktionen erwarten lassen. Oft wird eine vorgängige demographische Studie (Studie der Lebensgewohnheiten, der Arbeitszeiten usw.) für jeden Abonnenten-Typ nötig sein.

Es darf hier nicht vergessen werden, dass die Verwendung von grösseren geographischen Zonen die Homogenität erhöht.

2.133. Ein weiteres bei der Festlegung der physikalischen Grundlagen der Gleichungen zu berücksichtigender, wichtiger Gesichtspunkt ist das Kriterium der Linearität.

Dies gilt für das Gebiet der Analysen nach geographischen Zonen. Bei den Untersuchungen in der Zeit ist die Linearität im Grunde das Ziel der Analyse: man bestimmt zuerst die Beziehungen zwischen der Belastung und den verschiedenen Faktoren; auf Grund der Resultate stellt man alsdann fest, ob die a priori angenommene Linearität in Funktion der Temperatur oder des Logarithmus der Beleuchtungsstärke sich nachträglich bewahrheitet.

Auf dem Gebiet der Analysen nach geographischen Zonen, bei welchen lineare Beziehungen Voraussetzung sind, wird man danach trachten, durch die Wahl der physikalischen Grundlagen dafür zu sorgen, dass diese Hypothese a priori erfüllt ist. Wir lassen dabei die

Methode der «simultanen Gleichungen» ($X = 0$ oder 1; s. unter 1.132) beiseite, weil bei diesen von einer Linearität nicht mehr die Rede sein kann.

Es ist immer empfehlenswert, die physikalische Bedeutung der Regressionskoeffizienten b_1, $b_2 ... b_n$ zu unterstreichen.

Wenn z. B. die Analyse auf Grund von X-Werten durchgeführt wird, welche jährliche Verbrauchszahlen darstellen, während Y eine Belastung ist, so sind die Koeffizienten b die reziproken Werte der Benützungsdauer der den Verbrauchszahlen entsprechenden Belastungen. Eine genügende Linearität kann als gesichert betrachtet werden, wenn angenommen werden kann, dass die Benützungsdauer für die gleichartige Abonnentengruppe jeder Zone im Mittel die gleiche ist (im Grundehandelt es sich hier wiederum um das bereits früher erwähnte Homogenitätskriterium). Dies könnte dazu führen, in einer Abonnentengruppe verschiedene Untergruppen zu unterscheiden: s. Untersuchungsbeispiele (am Schluss des Berichtes): Studie Nr. 13, Unterscheidung zwischen Industrieabnehmern mit einer jährlichen Belastung über oder unter 1000 kW.

Weiteres Beispiel: Wenn die X Anschlusswerte sind, und Y immer noch eine Gesamtbelastung darstellt, so sind die b Belastungskoeffizienten (deren Wert zwischen 0 und 1 variiert), die den Anteil der betreffenden Anwendung an der Gesamtbelastung angeben. Der Belastungskoeffizient hat aber im allgemeinen sinkende Tendenz bei zunehmender Grösse des Abnehmers. Diese Schwierigkeit kann dadurch überwunden werden, dass die Analyse auf Grund einer geschichteten Stichprobe erfolgt, wobei die Schichten durch den Verbrauch der Abonnenten bestimmt werden (siehe Untersuchungsbeispiele am Schluss des Berichtes, Studie Nr. 15). Für jede Schicht wird dann eine Regressionsanalyse durchgeführt und die Ergebnisse werden mit entsprechender Gewichtung kombiniert.

Schliesslich sei noch bemerkt, dass bezüglich der Linearität wie auch der Homogenität die Resultate desto besser sein werden, je grösser die geographischen Zonen sind. Oft sind diese zwei Merkmale (Linearität und Homogenität) verknüpft; homogene Verbrauchergruppen verbessern die Linearität.

2.134. Am Schlusse dieser Betrachtungen über die Wahl der physikalischen Grundlagen der Gleichungen sei noch folgendes festgehalten:

— die geographischen Zonen müssen selbstverständlich so gewählt werden, dass der Variationsbereich der unabhängigen Variablen so weit als möglich ist, d. h. so, dass sie in bezug auf den jeweiligen Anteil der zu bestimmenden Grössen möglichst voneinander abweichen (im Grenzfall, wenn eine geographische Zone nur eine der Abonnentengruppen einschliesst, deren Belastung bestimmt werden muss, ist das Problem in bezug auf diese Belastung ohne Analyse bereits gelöst);
— die geographischen Zonen sollten praktisch von der gleichen Grössenordnung sein, weil sonst die Unterschiede zu einer Korrelation zwischen Abonnenten-Kategorien führen könnten (in der kleinen Zone erreichen die X nur niedrige Werte, während sich in den ausgedehnteren Zonen im allgemeinen höhere Werte annehmen);
— bei Untersuchungen mit Abonnentengruppen sollen diese Gruppen genügend gross sein, um das Spiel der Gleichzeitigkeit (Verschachtelung) zu gewährleisten.

2.14. d) *Wahl der Anzahl Gleichungen*

2.141. Die Zahl der Gleichungen soll selbstverständlich viel höher sein als die Zahl der Variablen, und zwar

— je grösser der Schwankungsbereich einer Variablen innerhalb der verfügbaren Angaben ist;
— je grösser die Ungenauigkeit der Angaben ist (vor allem wenn diese Angaben aus einer Stichprobe stammen);
— je grösser die Streuung in bezug auf die als linear angenommene Beziehung ist.

Auf jeden Fall wird unter sonst gleichen Bedingungen die Genauigkeit mit zunehmender Zahl der Gleichungen steigen.

Fortsetzung folgt

Fragen der Personenversicherungen in der Elektrizitätswirtschaft

Bericht über die 24. Diskussionsversammlung des VSE vom 23. November 1961 in Zürich und vom 8. März 1962 in Lausanne

Diskussion an der Versammlung in Lausanne

I. Kapitaldeckungsverfahren oder Umlageverfahren

Auch neuestens beschäftigt man sich in der Presse wieder mit dem Problem der Finanzierung von Pensionskassen: die einen stehen zum *Kapitaldeckungsverfahren*, die andern werben für das *Umlageverfahren*.

Ch. Maeder (Adm. générale des Services industriels de Genève):

Herr Dr. Zihlmann hat heute morgen darauf hingewiesen, dass nach einer gewissen Anzahl von Jahren die Umlageprämien ständig höher ausfallen würden als die dauernd gleichbleibenden Prämien des Dekkungsverfahrens. Nun hat meines Wissens die Stadt Lausanne trotzdem eine Finanzierungsmethode eingeführt, die auf dem Kapitaldeckungsverfahren *und* dem Umlageverfahren beruht (je 50 %); Vertreter des Personals und der Gewerkschaften in Genf sind ebenfalls der Ansicht, dass ein gemischtes System beträchtliche Vorteile bieten würde.

Dr. *E. Zihlmann* (CKW):

Versicherungsmathematische Berechnungen haben — wie gesagt — eindeutig ergeben, dass die Finanzierung mit Hilfe des Repartitionsverfahrens mit der Zeit wesentlich teurer zu stehen kommt; die höheren Prä-