

Stephan Wöbbeking

Design und Implementation eines
Softwaresystems für die Klassifikation und
Prognose von Zeitreihen

Diplomarbeit

BEI GRIN MACHT SICH IHR WISSEN BEZAHLT



- Wir veröffentlichen Ihre Hausarbeit, Bachelor- und Masterarbeit
- Ihr eigenes eBook und Buch - weltweit in allen wichtigen Shops
- Verdienen Sie an jedem Verkauf

Jetzt bei www.GRIN.com hochladen
und kostenlos publizieren



Wöbbeking, Stephan:
Design und Implementation eines Softwaresystems für die Klassifikation und
Prognose von Zeitreihen - 2001. - 148 S. Mittweida, Hochschule Mittweida (FH)
- University of Applied Sciences, Fachbereich MPI, Diplomarbeit, 2001

Referat:

Das Ziel der Diplomarbeit ist es, ein Vorhersagesystem für Zeitreihen zur Verfügung zu stellen. Es ist speziell ausgelegt auf die Bedürfnisse von Handel und Wirtschaft und berücksichtigt die Besonderheiten in den vorgelegten Datenbeständen. Das System untersucht die verschiedenen Merkmale in den Reihen und zeigt einige Lösungsansätze auf Basis einer Klassifikation auf. Diese Vorgehensweisen müssen in der Praxis getestet und weiterentwickelt werden. Die vorgelegte Implementierung zeigt einen Weg auf, wie mit dem System S-PLUS strukturiert programmiert werden kann. Es zeigt ebenfalls einen Weg für ein durchgängiges Datenkonzept auf. Die Arbeit umfaßt die drei wesentlichen Komponenten: ein Vorhersagesystem (Prognose), die Kennwertberechnung sowie die Visualisierung.

Inhaltsverzeichnis

1	Einführung	1
1.1	Aufgabe	2
1.2	Problemstellung	3
1.2.1	Prognose	3
1.2.2	Visualisierung	3
1.3	Rahmenbedingungen: Ein Praktikum bei der Firma Siemens	3
1.4	Überblick über die zeitliche Entwicklung	3
1.5	Die Wahl der Programmiersprache	5
1.6	Aufbau der Datenbasis	5
2	Grundlagen	7
2.1	Methoden der Regression	8
2.2	Die Modellbildung	8
2.3	Klassifikation	10
3	Problemanalyse und Lösungsansätze	12
3.1	Der Kennwert: volume weighted accuracy	13
3.2	Modelle	15
3.3	Notwendigkeit einer Klassifikation	17
3.4	Nullwerte	17
3.5	Weitere Modelle	19
3.6	Spitzenwerte	19
3.7	Unterteilung der Spitzenwerte	19
3.8	Der gleitende Durchschnitt	20
3.9	Trendwechsel	21
3.10	Negative Werte	25
3.11	Visualisierung	25
4	Systementwurf	26
4.1	Programmkomplexe und Namensgebung für Funktionen	27
4.2	Datenstruktur	28
4.2.1	Wertetabellen	28
4.2.2	Ergebnisdaten	28
4.3	Vorhersageberechnung	30

4.3.1	Vorbehandlung der Datenbasis	30
4.3.2	Gleitwerte	32
4.3.3	Kennwerte	33
4.3.4	Klassifikation	36
4.3.5	Modelle	38
4.3.6	Nachbearbeitung der Vorhersagen	38
4.4	Komplex Kennwertberechnung	38
4.4.1	Die Ergebniskomponenten	38
4.4.2	Aufbau des Komplexes	40
4.5	Visualisierung	40
4.6	Ergänzende Systemteile	42
5	Ergebnisse der Prognosen	43
5.1	Vergleich mit kommerziellen Produkten	44
5.1.1	Manugistics	44
5.1.2	AutoBox	44
5.1.3	Die Modelle L0 bis L8	44
5.1.4	Wertung	46
5.2	Vergleich von Mittelwert und Median	47
5.3	Vorhersagezeitraum	48
5.4	Eine erste Klassifikation	51
5.5	Weitere Modelle	53
5.6	Eine erste Modellauswahl	54
5.7	Große Spitzen	55
5.7.1	Modellrechnung für große Spitzen	55
5.7.2	Entfernen großer Spitzen	56
5.8	Trendwechsellpunkte	61
5.8.1	Effizienz der Trendwechsellpunkterkennung	61
5.8.2	Parameter für Trendwechsellpunkte	61
5.9	Vorhersage negativer Werte	62
5.10	Menge der bekannten Werte	64
5.11	Einfluß von Rundungen	69
5.12	Schranke für Nulltendenzwerte	69
5.13	PostHoc-Berechnungen	69
5.14	Rechenaufwand für die durchgeführten Berechnungen	70
5.15	Abschließende Wertung	72
6	Ausblick	74
6.1	Verbesserte change point Erkennung	75
6.2	Robustes Verhalten nach Trendwechsel	75
6.3	Andere „side-values“ für rMedian und lpVal	75
6.4	Klassifikation nur nach einem Wechsellpunkt	76
6.5	Andere Modelle integrieren	76
6.6	Andere Klassen für bestimmte Zeitreihen	76
6.7	Korrelation bei Zeitreihen einer Region	77

6.8	Integration weiterer Merkmale; Verbesserung der Klassifikation	77
6.9	Abweichungen der Modellwahl von einer PostHoc -Wahl	77
6.10	Implementation neuronaler Netze für Vorhersagen	78
6.11	Iteratives Eliminieren großer Spitzen	78

Anhang 80

A	Begriffserläuterung	80
A.1	Accuracy	81
A.2	ACM	81
A.3	AutoBox	81
A.4	changepoint	81
A.5	changepoint detection	81
A.6	classAve	82
A.7	classCount	82
A.8	classDistDia	82
A.9	gleitender Durchschnitt	82
A.10	große Spitze	82
A.11	itemAve	82
A.12	lag.x	83
A.13	lpVal	83
A.14	Manugistics	83
A.15	monthAve	83
A.16	null	83
A.17	ODBC	83
A.18	Oracle	84
A.19	params	84
A.20	peak	84
A.21	perClass	84
A.22	perItem	84
A.23	perMonth	85
A.24	PostHoc	85
A.25	Result	85
A.26	Siemens Corporate Research	85
A.27	Spitze	86
A.28	SSP - Siemens students program	86
A.29	Trendwechsel	86
A.30	volume weighted accuracy	86
B	Das Programmiersystem S-PLUS	87
B.1	Allgemeines	88
B.2	Syntax-Überblick	90
B.3	Datentypen	91
B.4	Arbeiten mit Vektoren	92
B.5	Arbeiten mit Matrizen, Arrays und Dataframes	94

B.6	Arbeiten mit Listen	97
B.7	Funktionsdefinitionen	98
B.8	Daten- und Programmstrukturen	99
B.9	Datenaustausch	100
C	Weitere Ergebnisse der Prognosen	102
D	Implementierung ausgewählter Details	111
D.1	Ergänzungen zur Datenstruktur	112
D.1.1	Verschiedene Wertetabellen	112
D.1.2	Verwaltung der Ergebnisse	112
D.2	Die drei Programmkomplexe	113
D.3	Protokollsystem	113
D.3.1	Bezeichner	113
D.3.2	Funktionsweise	114
D.3.3	Drucken einer Zeile	114
D.3.4	Ausgabe einzelner Zeichen	115
D.3.5	Initialisierung des Systems	117
D.3.6	Protokollklassen	117
D.3.7	Aktueller Ausgabelevel	118
D.3.8	Protokolldatei	118
D.3.9	Zeitangaben	118
D.3.10	Baumstruktur	119
D.4	Datumsangaben	119
D.4.1	Format des realDate	119
D.4.2	Scannen des textuellen Datums	119
D.4.3	Konvertierung in Textform	120
D.5	Prognosesystem	120
D.5.1	Vorbehandlung der Daten	120
D.5.2	Die Einsprungfunktion	123
D.5.3	Trendlinie	130
D.6	Kennwertberechnung	131
D.6.1	Die Funktion Accuracy	131
D.6.2	Die Funktion monthlyAccs	136
D.7	Diagrammgenerator	138
D.8	Die Hilfsfunktion paramsDiff	140
D.9	Die Hilfsfunktion listData	146
D.10	Die Hilfsfunktion listParams	146
	Literaturverzeichnis	148
	Selbstständigkeitserklärung	148

Tabellenverzeichnis

3.1	Der Mittelwert über Genauigkeiten	14
3.2	Bildung der volume weighted accuracy	14
3.3	Erste angewandte Modelle	16
4.1	Die Programmkomplexe	28
4.2	Ausschnitt aus der Datenbasis	29
4.3	Aufbau der Datenbasis	29
4.4	Angewendete Modelle	39
5.1	Manugistics -Ergebnisse	45
5.2	AutoBox -Ergebnisse	46
5.3	Die Modelle L0 bis L8	47
5.4	Vorhersagezeitpunkt / lag.x	49
5.5	Genauigkeiten der ersten 9 Modelle	51
5.6	Ergebnisse der „ <i>eyeball classification</i> “	52
5.7	Verwendete Klassifikation	54
5.8	Die Modellzuordnung	57
5.9	Monatserfolge nach Modellzuordnung	58
5.10	Vorhersagen mit Modellzuordnung	58
5.11	Versuch mit 20 Testreihen	59
5.12	Die Klassenergebnisse der Spitzeneliminierung	59
5.13	Die auf die Reihen bezogenen Ergebnisse der Spitzeneliminierung	60
5.14	Genauigkeit nach Monaten	62
5.15	Genauigkeit nach Klassen	63
5.16	Der Parameterraum über alle Klassen	63
5.17	Monatsergebnisse Negativvorhersagen	67
5.18	Klassenergebnisse Negativvorhersagen	67
5.19	Verschieden viele bekannte Werte	68
5.20	Rechenzeiten für Berechnung „bekannte Werte“	72
B.1	Basisdatentypen	91
B.2	Die erweiterten Datentypen	92
B.3	Spezielle Datentypen	92
C.1	Optionen für den Aufruf der Prognose	105

C.2	Zahlenwerte des Vergleichs Mittelwert - Median	106
C.3	Monatsergebnisse bei Rundung	107
C.4	Rundungen in Klassen	107
C.5	Auswirkung von Rundungen auf die Zeitreihen	108
C.6	Klassenauswertung Parameter Nulltendenz	108
C.7	Verteilung der Klassen bei veränderten Nulltendenzwerten	109
C.8	Möglichkeiten durch Modellwahl	109
C.9	Mögliche Klassengüten durch Modellwahl	110
D.1	Im System angelegte Datenextrakte	112
D.2	Die Programmkomplexe und einige Funktionen	113
D.3	Ausgabeklassen für das prot-System	118

Abbildungsverzeichnis

2.1	Regressionsarten	8
2.2	Summation zweier Kurven	9
2.3	Phasenverschiebung durch Koeffizienten	10
3.1	Ergebnisse der Modelle L0 bis L8 sowie erste PostHoc-Ergebnisse	16
3.2	Verschiedene Zeitreihen	18
3.3	Schranken für Spitzenidentifikation	21
3.4	Vergleich durch Mittelwert und Median gemittelte Zeitreihe . . .	22
3.5	Beispiel für einen Trendwechsellpunkt	23
3.6	Möglichkeiten der Positionierung für die Ausschnitte	23
3.7	Quotienten aus nebeneinanderliegenden Mittelwerten	24
4.1	Struktur für das Prognosesystem	31
4.2	Beispiel eines Spitzenwertes	32
4.3	Die Gleitwerte einer Zeitreihe	33
4.4	Schranken für Spitzen einer Zeitreihe	35
4.5	Zeitreihe aus der Klasse null	37
4.6	Zeitreihe aus der Klasse peak	37
4.7	Struktur des Komplexes Kennwertberechnung	40
4.8	Struktur des Diagrammgenerators	41
5.1	Vergleich Median und Mittelwert	48
5.2	Ähnliches Verhalten aller Modelle	50
5.3	Modelle und Klassen	55
5.4	Klassenverteilung	56
5.5	Genauigkeiten nach Monaten	64
5.6	Genauigkeiten nach Klassen	65
5.7	Parameterkombinationen	66
5.8	PostHoc-Berechnung	70
5.9	PostHoc-Berechnung über alle Klassen	71
5.10	Rechenzeitbeispiel	73
6.1	Einige Zeitreihen mit großer Übereinstimmung	79
B.1	Das S-PLUS System	88

D.1	Beispiel für Diagramm 1	141
D.2	Beispiel für Diagramm 2	141
D.3	Beispiel für Diagramm 3	142
D.4	Beispiel für Diagramm 4	142
D.5	Beispiel für Diagramm 5	143
D.6	Beispiel für Diagramm 6	143
D.7	Beispiel für Diagramm 7	147

Quellenausschnitte

B.1 Erzeugen von Zahlenfolgen	94
B.2 Rechnen in Folgen und Vergleiche	94
B.3 Multiplikation in Folgen	95
B.4 Erzeugen von Arrays	95
B.5 Setzen und Abfragen von Matrizen	96
B.6 Interpretation einer Matrix als Vektor	96
B.7 Arbeiten mit Listen	98
B.8 Ungewolltes Undefinieren belegter Bezeichner	99
B.9 Aufruf von Funktionen	99
D.1 Beispiel eines Protokolls	115
D.2 Verwendung der Funktion <code>realDate.parse</code>	120
D.3 Verwendung der Funktion <code>realDate.print</code>	120
D.4 Programmlauf von <code>listData()</code>	146

Listings

D.1	Die Funktion <code>prot.putc</code>	116
D.2	Die Funktion <code>preCompute.V15.201</code>	121
D.3	Die Funktion <code>doCall.V15.201</code>	126
D.4	Die Hilfsfunktion <code>trendLine</code>	130
D.5	Die Funktion <code>Accuracy</code>	133
D.6	Die Funktion <code>monthlyAccs</code>	136
D.7	Die Funktion <code>resItemDia.V17.100</code>	139
D.8	Die Hilfsfunktion <code>paramsDiff</code>	144

Kapitel 1

Einführung

1.1 Aufgabe

Im Rahmen dieser Diplomarbeit wird ein Problem der Klassifikation und Prognose von Zeitreihen bearbeitet. Es behandelt die Vorhersage von Verkaufszahlen. Ein führender amerikanischer Hersteller von Glühbirnen möchte Produktion und Vertrieb effizienter gestalten. Daher ist es von entscheidender Bedeutung, gute Prognosen über die eintretenden Verkaufsmengen erstellen zu können. Eine breite Auswahl unterschiedlicher Produkte führt zu einer großen Menge auftretender Daten. Pro Monat werden für jeden Artikel in verschiedenen Regionalbereichen die Bestellmengen aufsummiert und in Form von Zeitreihen in einer Datenbank abgelegt. Diese Daten dienen als Ausgangspunkt für die vorgenommenen Untersuchungen. In einer ersten Abstraktionsstufe werden die Verkaufszahlen von dem eigentlichen Produkt gelöst. Es handelt sich nun nur noch um einfache Zeitreihen, deren Verhalten untersucht, klassifiziert und prognostiziert werden soll. Diese Abstraktion wird bereits seitens des Herstellers vorgenommen und soll daher auch nicht im einzelnen erläutert werden. Die folgenden Untersuchungen beziehen sich daher auf Zeitreihen mit den verschiedenen Besonderheiten.

Es werden Wege und Ansätze aufgezeigt, die eher zu einer zufriedenstellenden Vorhersage der Zeitreihen führen können, als dies mit untersuchten Komplettlösungen möglich ist. So stand am Anfang die Beurteilung zweier verfügbarer Systeme, die laut Herstellerangaben bereits gute Ergebnisse liefern können. Nachdem sich jedoch herausgestellt hatte, daß diese Lösungen zu ungenauen Ergebnissen liefern, fiel die Entscheidung, eine eigene Programmierung vorzunehmen. Als Sprache stand hier das in Abschnitt 1.4 charakterisierte Paket S-PLUS zur Verfügung. Es waren bereits wenige Lösungsansätze erarbeitet worden. Diese ordneten sich allerdings nicht in ein Komplettsystem ein, sondern waren vielmehr unzusammenhängende Teilimplementationen für Ausschnitte aus dem Gesamtproblem. Sie mußten sowohl aufbereitet als auch um weitere Teillösungen erweitert werden. Der vorhandene Datenbestand war ungeordnet in verschiedenen Datensammlungen und Formaten verfügbar. Die Erarbeitung eines möglichst einfachen, jedoch universellen Datenkonzeptes wurde angestrebt. Anschließend wurden einige Programmfragmente analysiert, um die umgesetzten Algorithmen herauszufiltern und zu ordnen.

Die Problemanalyse schloß die Analyse der vorhandenen Ideen und die Überprüfung ihrer Effizienz ein. Nach einer ersten Datenanalyse wurde ein Konzept erarbeitet, um Strukturen und Relationen zu erkennen. Innerhalb dieses Systems ist es möglich, die erforderlichen Berechnungen durchzuführen sowie verschiedene Ergänzungen vorzunehmen. Hierbei wurde größtmöglicher Wert auf die Erweiterbarkeit, Wiederverwendbarkeit und Wartbarkeit gelegt, da weitere Untersuchungen durchgeführt werden sollen und diese zu weiteren Ergänzungen führen könnten. Die weitere Analyse der Daten soll mit Hilfe des erstellten Systems leichter durchgeführt werden können. Soweit möglich, werden sinnvolle Änderungen direkt in das Programmsystem übernommen.

Das Ziel dieser Diplomarbeit ist es, mathematische Modelle zu entwickeln und in ein Programmsystem umzusetzen. Dieses wird um einfache Berechnungen ergänzt, die die Modelle unterstützen und ergänzen, jedoch grundsätzlich nicht verändern. Der Schwerpunkt liegt somit in der Programmierung eines Systems, mit dessen Hilfe der Nutzer Prognosen erstellen, visualisieren und später auch deren Genauigkeit darstellen kann.

1.2 Problemstellung

Die gestellte Aufgabe läßt sich in zwei große Teilaufgaben gliedern. In einem ersten Teil sollen Vorhersagen für die gegebenen Zeitreihen gemacht werden. Anschließend ist nach einer geeigneten Möglichkeit zu suchen, dem Nutzer diese Daten zugänglich zu machen.

1.2.1 Prognose

Das weit größere der beiden Teilprobleme ist die Prognose der Zeitreihen. Laut der Aufgabenstellung sollen bei Nichtkenntnis der zukünftigen Werte diese möglichst genau prognostiziert werden. Bezüglich der Rechenzeit oder dem Einsatz professioneller Produkte usw. bestehen keine Einschränkungen.

1.2.2 Visualisierung

In einem zweiten Schritt ist die Bereitstellung einer geeigneten Präsentationsmöglichkeit gefordert. Die prognostizierten Werte sollten nicht nur als Zahlenwerte vorliegen, sondern optisch aufbereitet werden und mit dem vergangenen Verlauf der Zeitreihe zusammen visualisiert werden.

1.3 Rahmenbedingungen: Ein Praktikum bei der Firma Siemens

Die vorgelegte Arbeit wurde abgeleitet aus den Tätigkeiten, die im Rahmen eines Praktikums ausgeführt wurden. Dieses Praktikum wurde bei der Firma **Siemens Corporate Research** in Princeton / New Jersey in den Vereinigten Staaten durchgeführt. Es kam im Rahmen des Austauschprogramms **SSP - Siemens students program** zustande. Die Haupttätigkeiten in den USA waren einige Voruntersuchungen zu dem behandelten Thema, der Datenbankanbindung und der Analyse vorgegebener Lösungen.

1.4 Überblick über die zeitliche Entwicklung

Eine immer höhere Rechenleistung moderner Computer ermöglicht heute komplexe Berechnungen, die vor wenigen Jahren noch nicht durchführbar waren.

Ohne diese Entwicklung würde die Bewältigung einer gestellten Aufgabe so lange dauern, daß das erzielte Ergebnis schon nicht mehr von Bedeutung ist, wenn man denn überhaupt zu einem Ergebnis gekommen wäre. So werden viele Problemstellungen, für die bisher keine ausreichende Rechenleistung zur Verfügung stand, heutzutage neu aufgegriffen. Mit Hilfe dieser hohen Leistung werden Versuche unternommen, um die Berechnungen zu erweitern, zu verbessern oder überhaupt erst in einer angemessenen Zeit durchzuführen.

Auf den Bereich der Klassifikation von Zeitreihen zum Zwecke ihrer Einschätzung, Charakterisierung und deren Prognose trifft dies im besonderen zu. Zu dieser Thematik wird mit der vorgelegten Arbeit ein Beitrag geleistet. Es wird mit so umfangreichen Datenbeständen gearbeitet, daß eine Bearbeitung ohne den intensiven Einsatz der Datenverarbeitungstechnik unmöglich ist. Seit diese Verarbeitung jedoch möglich ist, werden Anstrengungen unternommen, neue Systeme zu schaffen und zu erweitern. Diese können umfangreiche statistische Kalkulationen durchführen. So werden manchmal Informationen gewonnen, deren Berechnung bisher nicht möglich war, die für die Lösung der verschiedensten Probleme jedoch hilfreich sein können.

Eine Reihe kommerzieller Produkte sind mittlerweile auf dem Markt, die darauf spezialisiert sind, die Hersteller- und Händlerbedürfnisse zu befriedigen. Es geht meistens darum, das Kundenverhalten besser einzuschätzen und Aussagen für die Zukunft treffen zu können. Zuverlässige Vorhersagen in diesem Zusammenhang sind von beträchtlichem Interesse, da sie in nahezu allen Branchen zu extremer Kosteneinsparung führen. Es sei insbesondere auf die Produktion, den Transport und die Lagerhaltung hingewiesen. Da jedoch nahezu jeder Datenbestand besondere Merkmale aufweist, liefern die untersuchten Komplettpakete keine zufriedenstellenden Lösungen. Mit ihnen wird der Versuch unternommen, die Probleme zu generalisieren, die besser individuell bearbeitet werden sollten.

Es bietet sich die Möglichkeit, jedes mathematische oder statistische Problem mittels einer allgemeinen Programmiersprache wie Pascal, C, Java etc. innerhalb eines ausführbaren Programmes zu lösen. Hierbei besteht jedoch die Problematik, daß diese Programmiersprachen nicht auf den Einsatz in der Mathematik spezialisiert sind und im Sprachstandard nur über sehr geringe mathematische Sprachelemente verfügen. Die Grundlagen sind implementiert, jedoch müssen alle weiterführenden Berechnungen vom Programmierer umgesetzt werden. Dies erfordert einen extrem hohen Aufwand an Personal und Zeit und verursacht dadurch enorme Kosten. Da gerade die Kosten- und Zeiteinsparungen die Gründe sind, warum solche Probleme gelöst werden sollen, ist dies ein unbefriedigendes Vorgehen.

Mittlerweile gibt es zusätzlich Produkte, die zwischen den Komplettlösungen und den Programmiersprachen einzuordnen sind. Vorwiegend handelt es sich hierbei um Programmiersprachen, denen herkömmliche Elemente fehlen, die dafür jedoch über sehr umfangreiche mathematische Funktionsbibliotheken ver-