4. Abschätzung der Prognosengüte

In diesem Kapitel sollen zunächst jene Methoden erläutert werden, welche zur Abschätzung der Leistungsfähigkeit der entwickelten statistischen Modelle verwendet wurden. Da diese bei nahezu allen Untersuchungsmethoden zum Einsatz kamen, wird in den folgenden Kapiteln nicht mehr explizit auf ihre Anwendung hingewiesen. Lediglich in den wenigen Ausnahmefällen, z.B. wenn die Untersuchungsmethode eine derartige Methodik nicht zuließ, werden diese Themen erneut näher behandelt. Die speziellen Details bei der jeweiligen Vorgehensweise werden hingegen erst im betreffenden Kapitel selbst erläutert.

4.1 Cross-Validation

Eine Gefahr bei der Korrelations- und Regressionsanalyse stellen die sogenannten Scheinkorrelationen dar. Da bei der Bestimmung der Korrelation nicht zwangsweise Größen verwendet werden müssen, die in einer physikalischen Beziehung zueinander stehen, kann es vorkommen, daß ein gefundener Zusammenhang lediglich von zufälliger Natur ist. So ist es ohne weiteres denkbar, daß sich bei der Berechnung der Korrelation zwischen den Schuhgrößen willkürlich ausgewählter Probanden und den täglichen Niederschlagsmengen an einer meteorologischen Station ein hoher Wert ergibt, obwohl augenscheinlich kein kausaler Zusammenhang zwischen beiden Größen besteht. Allgemein muß davon ausgegangen werden, daß der Korrelationskoeffizient einer endlichen Stichprobe immer einen von Null verschiedenen Wert annimmt. Die Gefahr einer Scheinkorrelation ist um so größer, je kleiner der Stichprobenumfang, d.h. die Objektanzahl, ist. Insbesondere dann können neben den interessierenden Gesetzmäßigkeiten auch die Besonderheiten, zumeist Zufälligkeiten der Stichprobe, modelliert werden (*Enke*, 1988). Die Erfahrung hat gezeigt, daß allein die Anwendung klassischer Signifikanztests nicht ausreicht, die Stabilität statistischer Beziehungen zu garantieren (*Enke*, 1988; *Balzer*, 1989). Diese Problematik verschärft sich noch, wenn eine Vielzahl potentieller Einflußgrößen (Prediktoren) zur Verfügung steht, da die Wahrscheinlichkeit zumindest eines positiven Testergebnisses (Überschreitung der Signifikanzschranke) mit jedem weiteren Prediktor immens anwächst. Testet man z.B. die Korrelationskoeffizienten von 20 potentiellen Prediktoren unter Verwendung klassischer Tests auf ihre Signifikanz hin, so besteht eine 64,2%ige Chance, daß zumindest ein Prediktor auf dem 5%-Niveau für signifikant erklärt wird. D.h. aber, daß lediglich mit 35,8%iger statistischer Sicherheit von einer Überzufälligkeit beim Auffinden genau eines „signifikanten“ Prediktors ausgegangen werden kann. Oft wird aber von der naiven Vorstellung ausgegangen, daß es sich um eine 95%ige Sicherheit handelt, da 5% von 20 exakt 1 ergibt (vergl. dazu Abschnitt 7.4). Dieses von *Wilks* (1995) als „multiplicity“ (dt. Vielfalt) bezeichnete Phänomen schwächt sich zwar mit sehr hoher Anzahl potentieller Prediktoren (ab etwa 1000) deutlich ab, ist jedoch auch dann von Bedeutung.

Speziell im Falle der multiplen linearen Regression kann es zum sogenannten Overfitting kommen. Hiervon spricht man, wenn die Vorhersageleistung eines statistischen Modells, welches an einem abhängigen Datenkollektiv (Entwicklungskollektiv) entwickelt wurde, aufgrund zu vieler verwendeter Prediktoren an einer unabhängigen Teilstichprobe (Examinationskollektiv) dramatisch einbricht. Eine Erklärung für dieses Phänomen liefert die Mathematik: Jede mögliche Kombination K=n-1 beliebiger Prediktoren liefert eine perfekte Anpassung an eine

maximal n Beobachtungen umfassende Stichprobe (jedoch mit unterschiedlichen Regressionskoeffizienten). Dies ist am einfachsten für den Fall n=2 zu veranschaulichen: Um die Regression den zwei Datenpunkten der Zielgröße anzupassen, bedarf es lediglich einer Geraden. Diese läßt sich durch nur einen Regressionskoeffizienten (Steigung) und den Achsenabschnitt vollständig beschreiben. Daher kann die Verwendung zu vieler, auch noch so sinnloser Prediktoren zu einer zufällig guten Anpassung im Entwicklungskollektiv führen. Beim Übergang zur unabhängigen Prognose erweisen sich diese Beziehungen jedoch als instabil, d.h. ihre Bestimmungsleistung läßt unweigerlich nach. Es ist leicht einzusehen, daß auch beim Overfitting sowohl die Objektanzahl als auch die Anzahl der potentiellen Prediktoren eine entscheidende Rolle spielen. Besondere Bedeutung kommt dabei dem Verhältnis dieser beiden Größen zu. Mit zunehmendem Stichprobenumfang steigt auch die Anzahl der Regressionskoeffizienten, die mit akzeptabler Genauigkeit geschätzt werden können (*Wilks*, 1995).
Eine geeignete Methode zur Beurteilung der Stabilität einer Regression stellt die sogenannte Cross-Validation dar (*Michaelson*, 1987). Ihre Verwendung ist insbesondere dann unerläßlich, wenn die Bereitstellung potentieller Prediktoren nicht auf physikalischer Basis möglich ist. Bei der Cross-Validation handelt es sich um eine spezielle resampling (dt. Neuanordnung) Technik. Dabei wird das Abschneiden des Modells wiederholt an einer Teilstichprobe getestet, die bei der Entwicklung der Regression zurückbehalten wurde. Bei einem n Beobachtungen umfassenden Datensatz wird also jeweils ein künstliches Verifikationskollektiv mit m Fällen geschaffen, mit dessen Hilfe die Leistungsfähigkeit der an den übrigen n-m Datenpunkten entwickelten Beziehung beurteilt werden kann. Theoretisch kann dieser Vorgang für alle (n!)/[(m!)(n-m)!] möglichen Aufteilungen durchlaufen werden. Häufig wird er jedoch nur so oft wiederholt, daß jede Beobachtung genau einmal dem Verifikationskollektiv angehörte. Üblich ist die Aufteilung der gesamten Stichprobe in zwei Hälften (H-Methode), vier Viertel (π-Methode) oder derart, daß das Entwicklungskollektiv jeweils n-1 Fälle enthält (U-Methode; auch Jackknife genannt).
Die Stabilität einer Regression kann unter Verwendung der Cross-Validation mittels zweier Kriterien eingeschätzt werden: Das Regularitätskriterium überprüft das Vorhersagepotential an unabhängigen Daten. In der Regel wird dabei ein skalares Fehlermaß (z.B. RMSE) verwendet, das jeweils für die m einzelnen Datenpunkte des künstlichen Verifikationskollektivs berechnet wird. Ist der komplette Zyklus dann durchlaufen, können diese unabhängigen Schätzungen noch zu einem einzelnen Gütemaß zusammengefaßt werden (z.B. RV-Wert). Eine andersartige Auswertung der durch die Cross-Validation produzierten Paare von Vorhersagen und Beobachtungen ist jedoch ebenfalls vorstellbar. In dieser Arbeit wurden sie teilweise zur Erstellung sogenannter „conditional quantile plots“, einer speziellen Form der „diagnostischen Verifikation“ (siehe Abschnitt 5.2), verwendet. Die verschiedenen Möglichkeiten zur Aufteilung der Stichprobe liefern leicht unterschiedliche Ergebnisse. „Während die H-Methode eine zu pessimistische Schätzung des Prognosefehlers auf Grund der drastischen Reduktion des Stichprobenumfanges liefert, gibt die U-Methode (theoretisch die optimale erwartungstreue Schätzung) keine Garantie für eine erwartungstreue Schätzung bei einer Modellbildung aus einer Vielzahl potentieller Prediktoren“ (*Enke*, 1988).
Das Kriterium der Unverzerrtheit untersucht die Variabilität der Regressionskoeffizienten, die an verschiedenen Teilen der Stichprobe entwickelt wurden. Je geringer diese ausfällt, desto mehr kann von einer hohen Stabilität des Modells ausgegangen werden. Im Falle der multiplen Regression wird zusätzlich die

Auswahl der Prediktoren bewertet. Eine geringe Übereinstimmung bei der Selektion der Prediktoren läßt auf instabile Verhältnisse schließen.
Die Methodik der Cross-Validation ist nicht ausschließlich auf die Anwendung im Rahmen von Regressionsanalysen beschränkt. Da sie parameterfrei ist, kann sie praktisch für jedes statistische Modell, bei dem eine Vorhersagevorschrift existiert, eingesetzt werden. Wird diese nicht aus dem Entwicklungskollektiv abgeleitet, sondern liegt bereits a priori vor, so kann die Unverzerrtheit des Modells natürlich nicht beurteilt werden. Die Abschätzung der Vorhersageleistung ist jedoch auch dann möglich. Ein Beispiel für einen solchen Fall stellen die Analogverfahren dar. Hier wird z.B. durch Mittelbildung der eingetroffenen Werte einer vorher festgelegten Anzahl analoger Fälle eine Prognose für die Zukunft erstellt. Entnimmt man dem historischen Archiv, aus dem die Analoga ausgewählt werden sollen, eine gewisse Anzahl von Fällen, so können diese als unabhängige Daten fungieren und das Abschneiden des Verfahrens kann getestet werden.

Einen hundertprozentigen Schutz vor Instabilität kann auch die Cross-Validation nicht garantieren. Das Risiko einer zufällig guten Anpassung bleibt erhalten; die Scheingüte (engl. artificial skill) wird durch ihre Anwendung lediglich minimiert. Insbesondere bei zu geringer Objektanzahl sowie dem Vorhandensein von (kurzfristigen) Trends bzw. Autokorrelationen ist Vorsicht angebracht (*Stephenson et al.*, 2000). Erst am vollständig unabhängigen Kollektiv, beim Übergang zur Vorhersagepraxis, stellt sich die tatsächliche Vorhersageleistung heraus. Bei ausreichend langen Zeitreihen sollte daher ein Teil der Stichprobe als Examinationskollektiv reserviert werden. Im Gegensatz zur Cross-Validation wird diese Teilstichprobe bei der Modellentwicklung überhaupt nicht mit einbezogen. Auf diese Weise kann die Vorhersageleistung noch effektiver abgeschätzt werden. Dabei sind „hindcasts“ möglichst zu vermeiden.
Da die für die Langfristprognose benötigten Datensätze jedoch zumeist nicht umfangreich genug sind, um ein Examinationskollektiv zu verwenden, muß der Fehler, mit dem die durch Cross-Validation abgeschätzte Prognosengüte behaftet ist, empirisch ermittelt werden. Die Verifikation operationeller Langfristvorhersagen bietet eine adäquate Möglichkeit.

4.2 Lineare Trendbereinigung

Werden statistische Vorhersagemodelle auf der Basis trendbehafteter Datenreihen entwickelt, so kann es zu einer Begünstigung der Prognosengüte kommen (vergl. z.B. *Barnston*, 1994). Diese kann sowohl zufälliger als auch kausaler Natur sein. Weist z.B. der Prediktand einer linearen Regression einen Trend auf, so wird jeder Prediktor, der mit einem gleichartigen oder entgegengesetzten Trend behaftet ist, eine hohe Korrelation zu diesem liefern. Setzen sich beide Trends im unabhängigen Kollektiv fort, so wird eine Prognosenleistung vorgetäuscht, sofern zwischen ihnen kein kausaler Zusammenhang besteht. Die Wahrscheinlichkeit eines zufällig übereinstimmenden Trends wächst natürlich mit steigender Anzahl potentieller Prediktoren. Da auch die Cross-Validation dazu neigt, die Prognosengüte bei Vorhandensein übereinstimmender Trends zu überschätzen (vergl. Abschnitt 4.1), wurden für diese Arbeit in der Regel sämtliche Datenreihen mittels (simpler) linearer Regression trendbereinigt. Dabei wurde darauf geachtet, daß die Mittelwerte der Zeitreihen erhalten blieben. Die Trendbereinigung wurde sowohl für die Prediktoren als auch für die Prediktanden durchgeführt. Das Risiko, daß in einigen Fällen Trends eliminiert wurden, zwischen

denen ein ursächlicher Zusammenhang besteht, wurde bewußt in Kauf genommen. Schließlich soll abgeschätzt werden, inwieweit die Leistungsfähigkeit der entwickelten Modelle über jene simpler Ansätze hinausreicht. Zur trivialen Trendvorhersage werden keine komplexen Modelle benötigt.
An dieser Stelle sei noch angemerkt, daß zusätzlich zur Trendbereinigung die Daten in die Form von Anomalien gebracht wurden. D.h. sämtliche Berechnungen beziehen sich auf die Mittelwerte der jeweiligen Reihen. Absolute Werte zukünftiger Vorhersagen müßten daher aus den betreffenden Steigungen und Mittelwerten rekonstruiert werden.