# 5. Verifikation

In diesem Kapitel sollen die in dieser Arbeit verwendeten Fehler- sowie Gütemaße kurz vorgestellt werden (Abschnitt 5.1). Zusätzlich zu diesen skalaren Maßen wurden im Rahmen einiger Untersuchungsmethoden sogenannte „conditional quantile plots“ erstellt. Sie stellen eine spezielle Form der „diagnostischen Verifikation“ dar und werden in Abschnitt 5.2 behandelt.

## 5.1 Skalare Maße

Um zu kontrollieren, ob ein systematischer Fehler vorliegt, wurde der „mean error (ME)“ oder Bias verwendet. Dieses Fehlermaß vergleicht ausschließlich den Mittelwert der Vorhersagen mit dem Mittelwert der Beobachtungen des Prediktanden:

$$ME = \frac{1}{n} \sum_{k=1}^{n} \left( o_k - y_k \right) = \bar{o} - \bar{y} \,. \tag{5.1}$$

Hierin ist $y_k$ die k. Vorhersage, $o_k$ die entsprechende Beobachtung. Ist der ME positiv, so sagt die Methode im Mittel zu kalt bzw. zu trocken vorher, fällt er negativ aus, so wurde zu warm bzw. zu naß vorhergesagt.

Der ME allein gibt noch keinen Aufschluß über die Genauigkeit einer Prognose. Ebenso wie z.B. Auflösung oder Schärfe ist auch der systematische Fehler lediglich ein Teilaspekt der Genauigkeit. Nur die Betrachtung sämtlicher zusammengehöriger Paare von Vorhersagen und Beobachtungen ermöglicht eine Aussage über die allgemeine Qualität der Prognose. Zu den am häufigsten verwendeten skalaren Genauigkeitsmaßen zählen der „mean absolute error (MAE)“

$$MAE = \frac{1}{n} \sum_{k=1}^{n} \left| o_k - y_k \right| \tag{5.2}$$

sowie der in Analogie zur Varianz definierte „mean-squared error (MSE)“:

$$MSE = \frac{1}{n} \sum_{k=1}^{n} \left( o_k - y_k \right)^2 \,. \tag{5.3}$$

Für beide Größen gilt, daß sie bei einer perfekten Prognose den Wert Null annehmen. Im Gegensatz zum linearen MAE reagiert der MSE auf große Fehler jedoch sehr sensibel. Sowohl der MAE als auch die Quadratwurzel des MSE, der sogenannte RMSE, stellen ein Maß für die typische Größenordnung des Vorhersagefehlers dar.

Neben den zu erwartenden Größenordnungen des systematischen und des allgemeinen Vorhersagefehlers ist vor allem die Leistungsfähigkeit eines Verfahrens von Interesse. Nur wenn dieses genauere Vorhersagen liefert als eine Standardprognose, ist es von Wert. Ein geeignetes Maß zur Beurteilung der Vorhersageleistung stellt die sogenannte „skill score“ dar. Sie ist ein Maß für die relative Genauigkeit der Vorhersagen in Bezug auf eine Referenzprognose. Die Art der Referenzprognose ist dabei im Prinzip beliebig, sie muß jedoch frei verfügbar und einfach zu bestimmen sein. Im allgemeinen kommen die Klimaprognose (Vorhersage des klimatologischen Mittelwertes), die Persistenzprognose (Erhaltungsneigung) oder eine Zufallsprognose zur Anwendung. Geben G und $G_{ref}$ die von einer Vorhersagemethode bzw. die von der Referenzprognose erzielte Genauigkeit an, so berechnet sich die „skill score (SS)“ wie folgt:

$$SS = \frac{G - G_{ref}}{G_{perf} - G_{ref}}. \tag{5.4}$$

Hierin ist $G_{perf}$ der Wert des verwendeten Genauigkeitsmaßes, den eine perfekte Prognose erzielen würde. Multipliziert man die SS mit dem Faktor 100, so ergibt sich die prozentuale Verbesserung durch das Vorhersagemodell.
Da die Klimavorhersage der Persistenzprognose im Bereich der für langfristige Witterungsvorhersagen relevanten Zeitskala deutlich überlegen ist, wurde sie zur Bestimmung der Vorhersageleistung verwendet. Als Genauigkeitsmaß kam in nahezu allen Fällen der MSE zur Anwendung. Somit vereinfacht sich Gleichung 5.4 zu:

$$SS = 1 - \frac{MSE}{MSE_{Klima}}. \tag{5.5}$$

Liegt die SS in dieser Form vor, so wird sie auch als Reduktion der Varianz (RV) bezeichnet. Der Wert $MSE_{Klima}$ entspricht dabei der Varianz des Prediktanden im Entwicklungszeitraum, da man von der Repräsentativität der verwendeten Stichprobe für die Grundgesamtheit auszugehen hat (vergl. *Röder*, 2001). Speziell in dieser Form gibt der RV daher Aufschluß darüber, um welchen Anteil die natürliche Schwankungsbreite der vorherzusagenden Größe durch die Vorhersagemethode reduziert werden kann. Nimmt er negative Werte an, so ist das Verfahren schlechter als die Klimavorhersage. Ebenso wie die SS kann auch die Reduktion der Varianz mit anderen Referenzprognosen bestimmt werden.

## 5.2 „Diagnostische Verifikation“ – „Conditional Quantile Plots“

Die große Stärke skalarer Genauigkeitsmaße, nämlich die Zusammenfassung der Vorhersagequalität in nur einer Größe, stellt gleichzeitig auch eine Schwäche dar. Eine derartige Vereinigung verschiedener Teilaspekte der Vorhersagequalität kann dazu führen, daß sich für mehrere Vorhersagestichproben mit unterschiedlicher Fehlercharakteristik dennoch ein und derselbe Wert eines bestimmten

Genauigkeitsmaßes ergibt. Eine Differenzierung dieser Stichproben wäre retrospektiv nicht mehr möglich. Speziell dann, wenn an Hand skalarer Genauigkeits- bzw. Gütemaße lediglich eine marginale Vorhersageleistung erzielt wird, kann deren ausschließliche Verwendung sogar zu einer zu pessimistischen Einschätzung des wissenschaftlichen Wertes und des potentiellen wirtschaftlichen Nutzens eines Vorhersagemodells führen (*Wilks*, 2000a).
Eine geeignete Alternative stellt die sogenannte diagnostische Verifikation dar. Sie beschreibt die statistischen Beziehungen der vollständigen gemeinsamen Verteilung der Vorhersagen und Beobachtungen in zusammenfassender Manier, wodurch die Identifikation spezieller Stärken und Schwächen in Bezug auf die verschiedenen Teilaspekte der Vorhersagequalität ermöglicht wird. Somit erlaubt sie den Verantwortlichen, gezielte Anstrengungen zur Verbesserung der Vorhersagen vorzunehmen und den Verbrauchern, sie optimal zu verwenden.

Da sowohl die vorliegenden Vorhersagen als auch die dazugehörigen Beobachtungen in der Praxis eine Vielzahl verschiedener Werte annehmen (die Anzahl möglicher Versuchsausgänge der kontinuierlichen Größen wird durch Rundung auf eine endliche Zahl begrenzt; $f_i$, i=1,...,I und $o_j$, j=1,...,J), mußte ihre vollständige gemeinsame Verteilung,

$$p(f_i, o_j) = \Pr\{f_i \cap o_j\}, \tag{5.6}$$

zwecks Übersichtlichkeit faktorisiert werden. Es sei darauf hingewiesen, daß die Unterscheidung zwischen der Wahrscheinlichkeitsdichte der Grundgesamtheit (rechte Seite) und den relativen Häufigkeiten der Stichprobe (linke Seite) in Gleichung 5.6 außer Acht gelassen wurde. *Wilks* (2000a) zufolge eignen sich insbesondere häufig sehr aufschlußreiche graphische Faktorisierungen der Form

$$p(f_i, o_j) = q(o_j \mid f_i)\ r(f_i), \tag{5.7}$$

die man als „calibration-refinement factorization“ bezeichnet, für eine detaillierte, aber dennoch überschaubare Verifikation. Diese drücken die gemeinsame Verteilung der Vorhersagen und Beobachtungen als das Produkt der bedingten Verteilungen der Beobachtungen für jede der I verschiedenen Vorhersagen $q(o_j \mid f_i)$ und der vollständigen (unbedingten) Häufigkeitsverteilung der Vorhersagen $r(f_i)$ aus. Gilt es, kategorische Vorhersagen eines kontinuierlichen Parameters zu verifizieren, so empfiehlt es sich, speziell sogenannte conditional quantile plots zu verwenden (*Wilks*, 1995).
Ein Beispiel für einen „conditional quantile plot“ findet man in Abbildung 5.1. Dargestellt sind die 1-Monats-Prognosen der monatlichen Temperaturanomalien, wie sie sich durch eine multiple lineare Regression (Kapitel 7) für ein 30-jähriges Examinationskollektiv ergeben haben. Zur Veranschaulichung sind Vorhersagen und Beobachtungen in übereinstimmende Klassen gleicher Breite eingeteilt. Es handelt sich daher bei den Werten $f_i$ und $o_j$ um Klassenmittel, wobei gilt: I=J und $f_i=o_i$. Zusätzlich wurden sowohl Vorhersagen als auch Beobachtungen standardisiert, um die Vergleichbarkeit der Ergebnisse sämtlicher Kalendermonate zu gewährleisten.

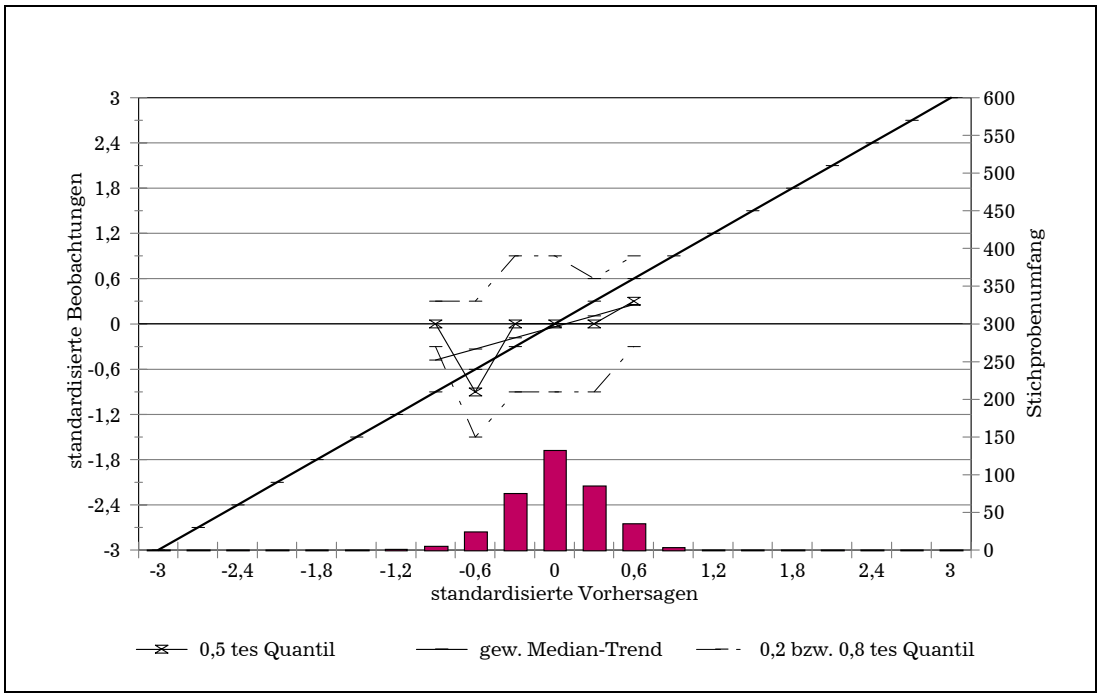


**Abbildung 5.1:** „Conditional quantile plot“ der 1-Monats-Prognosen der monatlichen Temperaturanomalien nach einer multiplen linearen Regression für den Zeitraum 1969 bis 1998.

Wie sich Abbildung 5.1 entnehmen läßt, enthält ein derartiges Diagramm zwei Teile, welche die beiden Komponenten der Gleichung 5.7 (rechte Seite) repräsentieren. Die bedingten Verteilungen der Beobachtungen bei festgehaltener Vorhersage werden in Form ausgesuchter Quantile dargestellt (linke y-Achse). Sie sind mit der 1:1-Diagonalen zu vergleichen, welche die perfekte Prognose repräsentiert. Das Histogramm im unteren Teil der Abbildung gibt Aufschluß über die absoluten Häufigkeiten der jeweiligen Vorhersagen (rechte y-Achse). Da die Klassen in der Regel sehr ungleich besetzt sind, wurde zur Vereinfachung der Güteeinschätzung ein gewichteter Trend des 0,5ten Quantils bestimmt. Im Folgenden wird daher vornehmlich dieser Median-Trend zur Bewertung herangezogen werden, um Fehleinschätzungen resultierend aus der Überbewertung gering besetzter Klassen zu vermeiden. Zur vollständigen Beurteilung der Vorhersagequalität müssen jedoch sämtliche Quantile betrachtet werden. Nur wenn diese einheitlich annähernd parallel zur Diagonalen verlaufen, sind tatsächlich alle Prognosen gut kalibriert. Des weiteren sind auch die Abstände der einzelnen Quantile zur Diagonalen von Bedeutung. Sie geben Aufschluß über die zu erwartende Streuung des Vorhersagefehlers. Für Klassen mit einem Stichprobenumfang unter fünf wurden keine bedingten Verteilungen der Beobachtungen erstellt, da sich bei so geringer Anzahl keine sinnvollen Quantile bestimmen lassen.

Insgesamt vier Teilaspekte oder Attribute der Vorhersagequalität können einem „conditional quantile plot“ entnommen werden:

Ein *Bias* liegt dann vor, wenn zumindest die Mehrzahl der Mediane der bedingten Verteilungen beständig oberhalb bzw. unterhalb der Diagonalen liegen.

Die *Schärfe* bezieht sich ausschließlich auf die Vorhersagen; sie charakterisiert ihre unbedingten Häufigkeitsverteilungen. Dabei weisen Vorhersagen, die nur selten und zumeist wenig vom klimatologischen Mittel abweichen, eine geringe Schärfe auf. Vorhersagen, die häufig beträchtlich vom Mittelwert abweichen, sind hingegen scharf.

Die *Verläßlichkeit* (engl. *reliability*) ist ein Maß dafür, inwieweit die bedingten mittleren Beobachtungen mit den jeweilig dazugehörigen Vorhersagen übereinstimmen. Je kleiner die Summe der Differenzen zwischen den Vorhersagen $f_i$ und den Medianen der entsprechenden Verteilungen ist, desto größer ist die Verläßlichkeit. Dieses Fehlermaß wird oft auch als eine Art zusammenfassende Größe für die Gesamtheit der I bedingten Verteilungen der „calibration-refinement factorization“ bezeichnet. Daher kann die Verläßlichkeit allein durch Betrachtung des Median-Trends korrekt eingeschätzt werden. Beträgt die Steigung der Geraden den Wert 1, so liegt die optimale Verläßlichkeit vor, sofern kein Bias vorhanden ist. Häufig wird der systematische Fehler als eigenständiger Teilaspekt bei der Beurteilung der Verläßlichkeit nicht mit berücksichtigt. Je mehr die Steigung vom Wert 1 abweicht, desto weniger verläßlich sind die Prognosen. Dabei sind zwei Fälle zu unterscheiden: Ist die Steigung deutlich größer als 1, so ist mehr Zuversicht seitens der Vorhersagenden angemessen, d.h. schärfere Prognosen würden zu einer Annäherung an die perfekte Prognose führen. Im „conditional quantile plot“ entspräche dies einer Drehung des Median-Trends nach rechts. Im Gegensatz dazu spiegelt eine Steigung zwischen 0 und 1 zu „mutige“ Vorhersagen wider. In diesem Falle würden konservativere Prognosen zu einer gesteigerten Verläßlichkeit beitragen (Drehung nach links).
Die *Auflösung* mißt den Unterschied zwischen den Werten der einzelnen Mediane. Je größer diese Differenz ist, desto mehr sind die Vorhersagen in der Lage, die verschiedenen Versuchsausgänge (Beobachtungen) aufzulösen. Liegen sämtliche Werte der Mediane in der gleichen Größenordnung (im vorliegenden Beispiel Klasse), d.h. ist der mittlere Versuchsausgang unabhängig vom Vorhersagewert, so existiert keine Auflösung. Auch dieses Maß läßt sich unmittelbar an Hand des Median-Trends abschätzen. Je größer die Steigung, desto höher die Auflösung. Rein arithmetisch liegt auch dann eine Auflösung vor, wenn die Steigung negative Werte annimmt. Derartige Vorhersagen weisen jedoch eine extrem schlechte Verläßlichkeit auf. Auch praktisch sind sie selbstverständlich nicht von Nutzen.