Informationstheorie bei der Evaluation von Vorhersagen

J. Tödter (toedter@iau.uni-frankfurt.de)[1] und B. Ahrens[1]
[1] Institut für Atmosphäre und Umwelt, Goethe-Universität, Frankfurt am Main

Motivation

Zur Evaluation probabilistischer Vorhersagen werden oft quadratische Fehlermaße verwendet. Vorhersagen sind inherent mit *Unsicherheit* und dem Vermitteln von *Informationen* verknüpft, daher bietet die Informationstheorie hier den natürlicheren Rahmen.

Grundlagen

O: Beobachtung, J Möglichkeiten (diskrete Kategorien)
F: Vorhersage, mit I möglichen Wahrscheinlichkeitsverteilungen $\mathbf{f_i}$
H: Entropie einer Wahrscheinlichkeitsverteilung $H(X) = -\sum p(x_i) \log p(x_i)$

Alle Informationen zur Verifikation sind enthalten in der Gemeinsamen Wahrscheinlichkeitsverteilung von F und O

Marginale Verteilung: Wie oft wurde jede bestimmte Vorhersage $\mathbf{f_i}$ verwendet?

$$p(\boldsymbol{f_i}, o_j) = p(\boldsymbol{f_i}) \cdot p(o_j|\boldsymbol{f_i})$$

Bedingte Verteilungen: Relative Häufigkeiten der Beobachtung nach einer bestimmten Vorhersage $\mathbf{f_i}$.

Bewertung durch Scores

Die Qualität (Genauigkeit) wird durch Scores gemessen. Die vorhergesagte Wahrscheinlichkeit *f* für die *realisierte Kategorie* sollte möglichst groß sein. Wir vergleichen hier:

informationstheoretische Größen & **quadratisches Fehlermaß**

Ignorance-Score: [3] Informationsgehalt in der Realisierung

$$IGN = -\log(f)$$

Brier-Score: Quadrierter Abstand bei der Realisierung

$$BS = (1-f)^2$$

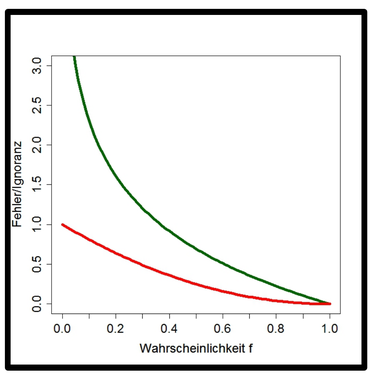


Fig.1: Fehler bzw. Informationsgehalt

$$\left.\begin{matrix} IGN \\ BS \end{matrix}\right\} = REL - RES + UNC$$

Durch analoge **Zerlegungen** der Scores erhält man weitere Attribute der Qualität und so eine detaillierte Auswertung:

REL - "Reliability" (bedingter Bias):

Die Verteilung der Beobachtung nach einer Vorhersage $\mathbf{f_i}$ sollte *im Mittel* dieser Vorhersageverteilung $\mathbf{f_i}$=f_{ij} entsprechen.

Mittlere **Relative Entropie**

$$\sum_i p(\boldsymbol{f_i}) \sum_j p(o_j|\boldsymbol{f_i}) \log \left[\frac{p(o_j|\boldsymbol{f_i})}{f_{ij}}\right]$$

$$\sum_i p(\boldsymbol{f_i}) \sum_j [p(o_j|\boldsymbol{f_i}) - f_{ij}]^2$$

Mittlere Quadratische Abweichung

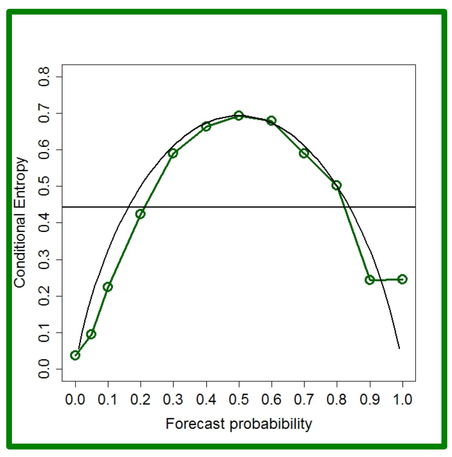
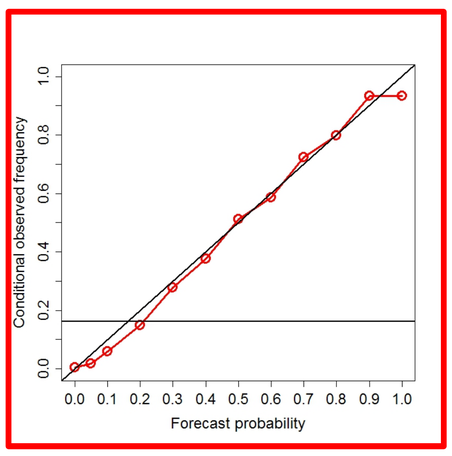


Fig.2+3: Informationstheoretisches Analogon (bedingte Entropie) im Vergleich zum üblichen "Reliability Diagram" [4]



RES - "Resolution":

Wie gut kann die Vorhersage zwischen verschiedenen nachfolgenden Beobachtungen unterscheiden?

Mittlere **Relative Entropie**

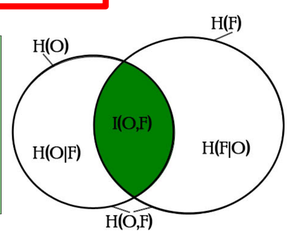
$$\sum_i p(\boldsymbol{f_i}) \sum_j p(o_j|\boldsymbol{f_i}) \log \left[\frac{p(o_j|\boldsymbol{f_i})}{p(o_j)}\right]$$

$$\sum_i p(\boldsymbol{f_i}) \sum_j [p(o_j|\boldsymbol{f_i}) - p(o_j)]^2$$

Mittlere Quadratische Abweichung

$$I = \sum_i p(\boldsymbol{f_i}) [H(O) - H(O|\boldsymbol{f_i})]$$

Transinformation: *Reduzierung der Unsicherheit aufgrund Kenntnis der Vorhersage - Wieviel besser ist die Vorhersage als eine unabhängige?*



UNC - "Uncertainty":

Maß für die Variabilität der Beobachtung und damit die Schwierigkeit der Vorhersage. Hängt nur von der Klimatologie ab.

Entropie H(O) der Beobachtung

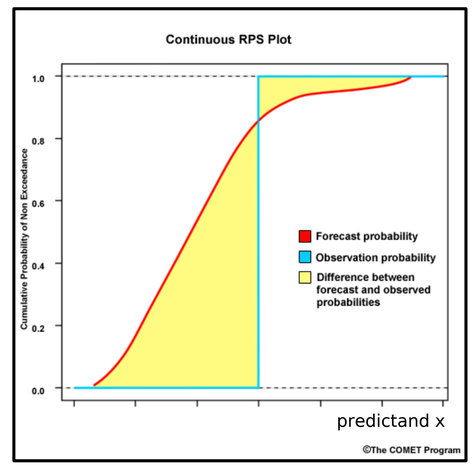
$$-\sum_j p(o_j) \log p(o_j)$$

$$\sum_j p(o_j)(1 - p(o_j))$$

Varianz der Beobachtung

Mehrere Kategorien: Ranked Scores

Für mehr als 2 Kategorien sollte bei ordinalen Größen die Verteilung der Wahrscheinlichkeiten in den nicht-realisierten Kategorien (***Distanz***) berücksichtigt werden.[1]
Dazu werden BS und analog IGN zu "**Ranked Scores**" weiterentwickelt:



➡ *Vergleich der kumulierten Verteilungen von Vorhersage und Beobachtung*

Das ist äquivalent zur Auswertung des IGN / BS für jeden Schwellwert als binäres Ereignis und anschließende Summierung über alle Klassen.[2] Analoges gilt für die Zerlegungen.

Kontinuierliche Generalisierung

Die (Ranked) Scores hängen von der Anzahl und Wahl der Schwellwerte zur Klasseneinteilung ab. Um das zu umgehen, wird im Falle des CRPS über alle möglichen Schwellwerte x integriert:
Dies ist analog auch für den IGN-Score möglich.

Gemeinsame Verteilung für jeden Schwellwert x

IGN(x) & Zerlegung(x) → $CRIGN = \int IGN(x)dx$

BS(x) & Zerlegung(x) → $CRPS = \int BS(x)dx$

Für Ensemble-generierte Vorhersagen geht die Integration über eine endliche Zahl von Schwellwerten und ist exakt durchführbar.

Näherung der Informationstheorie

Informationstheorie ist die **allgemeinere** Basis.
Es handelt sich nicht nur um eine formale Analogie, sondern auch um eine direkte mathematische Abhängigkeit:

Informationstheorie —*Taylorapprox. 2. Ordnung*→ Quadratische Fehlermaße

Beispiel

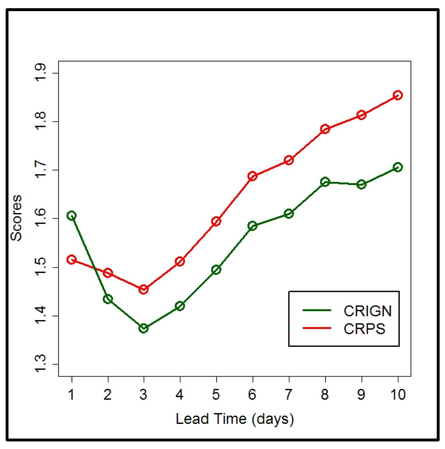
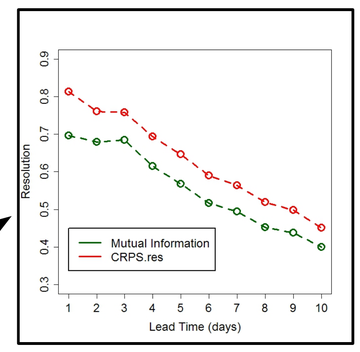


Fig 5+6.: Vergleich der Ranked Scores für Niederschlagsvorhersagen durch ein Ensemblesystem [5] in Abhängigkeit der Lead Time

Gezeigt wird die Abnahme der Qualität mit der Lead Time, die v.a. am Verlust von Resolution oder *Transinformation* liegt.



Die oben begründete Ähnlichkeit beider Maße ist deutlich sichtbar.

Fazit

- Informationstheorie bietet einen **natürlichen Rahmen** für die Evaluation probabilistischer Vorhersagen
  - Quadratische Fehlermaße stellen lediglich eine Näherung dar
  - Unterschiede ergeben sich vor allem für seltene Ereignisse
  - Berücksichtigung der Distanz und kontinuierliche Version möglich
- Neue Interpretation: Informationsgewinn und Reduktion der Unsicherheit für Nutzer statt "Fehler"

Referenzen

[1] Ahrens,B. and A.Walser (2008): Information-Based Skill Scores for Probabilistic Forecasts, Mon. Wea. Rev., 136, 352-363
[2] Hersbach,H. (2000): Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Wea. & Forecasting, 15, 559-570
[3] Roulston,M. and L.Smith (2002): Evaluating probabilistic forecasts using information theory. Mon. Wea. Rev., 130, 1653–1660.
[4] Beispieldaten aus: Wilks, "Statistical Methods in Atmospheric Sciences", 2006, p.270
[5] Beispieldatensatz aus dem R-"Verification"-Package, 2010