

Revenue Operations & KI

Was ist Zeitreihenprognose?

Zeitreihenprognose schreibt Reihen aus ihrer Vergangenheit fort. Im M4-Wettbewerb führten ein Hybrid und Kombinationen; reine Verfahren fielen zurück.

Autor	Veröffentlicht	Lesezeit	Wörter
Sinan Isoglu, MBA (Quantic)	18. Oktober 2026	20 Min.	4.241
Führungskraft für Umsatzwachstum, Dozent und Doktorand			

Management Summary

Zeitreihenprognose sagt die künftigen Werte einer Reihe aus deren eigener Vergangenheit vorher, indem sie Niveau, Trend und Saisonmuster fortschreibt. Dieser Konzeptleitfaden definiert sie, stellt die Verfahrensfamilien vor (naive und saisonale naive Benchmarks, exponentielle Glättung, ARIMA, Kombinationen, maschinelles Lernen und hybride Verfahren), berichtet, was der M4-Wettbewerb fand, als 61 Verfahren 100.000 Reihen prognostizierten, und erklärt, warum Hyndman und Koehler für den Vergleich der Genauigkeit über Reihen hinweg skalierte Fehler wie den mittleren absoluten skalierten Fehler vorschlagen statt Fehler in Einheiten oder Prozent. Die Quellen sind ein Prognosewettbewerb mit kontinuierlichen Reihen, der sporadische Nachfrage ausließ, und ein methodischer Aufsatz zu Fehlermaßen.

Inhalt	§ 1	Was ist Zeitreihenprognose?	2
	§ 2	Welche Verfahren schreiben eine Reihe aus ihrer Vergangenheit fort?	3
	§ 3	Was zeigten 100.000 Reihen über diese Verfahren?	5
	Abbildung 1	Genauigkeit von 14 der 61 M4-Verfahren, relativ zu Naïve 2	6
	§ 4	Warum lassen sich Prognosefehler nicht ohne Weiteres über Reihen mitteln?	9
	§ 5	Was sollte ein Prognosevergleich festlegen, bevor der erste Wert berechnet wird?	11
	Abbildung 2	Ein Blatt für einen Prognosevergleich	12
	§ 6	Welche dieser Ergebnisse hielten einer kurzen oder sporadischen Reihe nicht stand?	13
		Referenzen	15

Was ist Zeitreihenprognose?

§ 1 von 6

Sie prognostiziert eine Reihe aus ihrer eigenen Vergangenheit. Hyndman und Koehler (2006)¹ vergleichen nach eigener Angabe Maße für die Genauigkeit univariater Zeitreihenprognosen („discuss and compare measures of accuracy of univariate time series forecasts“), also von Prognosen für eine einzelne Reihe, und ihre vier Beispielverfahren, vom historischen Mittelwert („the historical mean using data up to the most recent observation“) bis zum Verfahren von Holt, verwenden nichts als die Reihe selbst. Im M4-Prognosewettbewerb gaben Makridakis et al. (2020)² den Teilnehmenden die Historie jeder Reihe und hielten ihr Ende zurück („The training set was made available to the participants at the beginning of the competition, while the test set was kept secret till its end“). Vor der Freigabe wurde alles entfernt, was die ursprüngliche Reihe hätte erkennbar machen können, auch das Startdatum („any information that could possibly lead to the identification of the original series was removed so as to ensure the objectivity of the results“). Als offene Frage nennen die Autoren, welche Verbesserung bei Punktprognosen und Prognoseintervallen möglich wäre, wenn die Zeitreihenprognose erklärende Variablen einbezöge („expanding time series forecasting to include explanatory/exogenous variables“).

¹ Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.

² Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74.

Makridakis et al.² beschreiben ihre Benchmark-Verfahren danach, welchen Teil des Musters einer Reihe sie fortschreiben: Naïve 2 erfasst nur die Saisonalität („captures seasonality only“), die einfache exponentielle Glättung das Niveau („captures the level“), das Verfahren von Holt schreibt einen linearen Trend fort („extrapolates using a linear trend“), und das gedämpfte Verfahren dämpft diesen Trend („dampens the linear trend“). In den Worten dieses Leitfadens ist das Niveau der Stand, auf dem die Reihe gerade liegt, der Trend die Richtung, in die sie sich bewegt, und die Saisonalität ein Muster, das sich über einen festen Zeitraum wiederholt, etwa über die Monate eines Jahres.

In diesem Leitfaden ist eine Zeitreihenprognose eine Prognose künftiger Werte einer Reihe, die aus den vergangenen Werten dieser Reihe entsteht, mit einem Verfahren, das angibt, wie es Niveau, Trend

und ein etwaiges Saisonmuster fortschreibt, und die an Werten gemessen wird, die das Verfahren nicht gesehen hat. Das Verfahren darf auch aus anderen Reihen derselben Art lernen: Die drei genauesten Verfahren in M4 zogen Informationen aus mehreren, nach Frequenz zusammengefassten Reihen heran („introduced information from multiple series (aggregated by data frequency)“), um für jede Reihe das Verfahren oder die Gewichte einer Kombination zu wählen (S. 14 der vorliegenden Vorabfassung [Article in Press]). Eine Prognose, die zusätzlich andere Variablen nutzt, etwa Preise, Aktionen oder die offenen Opportunities einer Vertriebspipeline, geht über diese Definition hinaus; M4 ließ solche Variablen außen vor, und dieser Leitfaden tut es auch. Die Beispiele hier sind Umsatz- und Nachfragerreihen. Die Datenbank, aus der M4 schöpfte, legt den Schwerpunkt auf betriebliche Prognoseanwendungen („ForeDeCk emphasizes business forecasting applications“), und die 100.000 Reihen stammen aus sechs Bereichen: Mikro, Industrie, Makro, Finanzen, Demografie und Sonstiges (Tabelle 1, S. 3 der Vorabfassung).

Welche Verfahren schreiben eine Reihe aus ihrer Vergangenheit fort?

§ 2 von 6

Am einfachsten sind die naiven Benchmarks. In M4 ist Naïve 1 ein Random-Walk-Modell, das annimmt, dass künftige Werte dem letzten bekannten Wert gleichen („A random walk model, assuming that future values will be the same as that of the last known observation.“). Naïve S, die saisonale naive Prognose, setzt jeden Prognosewert gleich dem letzten bekannten Wert derselben Periode („the last known observation of the same period“): Der nächste März wird mit dem letzten März prognostiziert. Naïve 2 entspricht Naïve 1, nur wird die Reihe vorher bei Bedarf mit einer klassischen multiplikativen Zerlegung saisonbereinigt („Like Naïve 1 but the data are seasonally adjusted, if needed, by applying a classical multiplicative decomposition“); ob eine Reihe saisonal ist, entscheidet ein Autokorrelationstest, den die Autoren als 90-Prozent-Test angeben. Nach der Lesart dieses Leitfadens sind sie Maßstäbe, keine Empfehlungen: Ein komplexeres Verfahren lohnt seinen Aufwand nur, wenn es sie an Daten schlägt, die es nicht gesehen hat.

Die exponentielle Glättung ist eine Familie. Die einfache exponentielle Glättung (SES) glättet die Daten exponentiell und schreibt sie ohne Trend fort („Exponentially smoothing the data and extrapolating assuming no trend.“); das Verfahren von Holt unterstellt einen linearen Trend, das gedämpfte Verfahren (Damped) einen gedämpften („a damped trend“); bei allen dreien wird in M4 wie bei Naïve 2

saisonbereinigt („Seasonal adjustments are considered as per Naïve 2.“). ETS wählt unter den Modellen der exponentiellen Glättung automatisch das beste nach Informationskriterien („Automatically provides the best exponential smoothing model, indicated through information criteria.“). Auch die Theta-Methode, das genaueste Verfahren des früheren M3-Wettbewerbs („the most accurate method of the M3 Competition“), ist ein Einzelverfahren; Hyndman und Koehler (2006)¹ verweisen auf einen Nachweis, dass sie einer einfachen exponentiellen Glättung mit Drift entspricht, deren Drift die halbe Steigung einer an die Daten angepassten linearen Regression ist („is equivalent to simple exponential smoothing (SES) with drift where the drift is half the value of the slope of a linear regression fitted to the data“).

ARIMA-Modelle stammen aus der Box-Jenkins-Methodik, die nach der Darstellung von Makridakis et al.² davon ausging, dass es für jede Zeitreihe ein optimales Modell gibt („claimed that an optimal model existed for each time series“), das sich durch Identifizieren, Schätzen und Prüfen finden lässt. In M4 entstand die ARIMA-Prognose automatisch: Aus den möglichen ARIMA-Modellen wird nach geeigneten Kriterien automatisch das beste gewählt („An automatic selection of possible ARIMA models is performed and the best one is chosen using appropriate selection criteria.“). Der M4-Bericht beschreibt ARIMA danach, wie ein Modell gewählt wird, nicht über seine Gleichungen, und dieser Leitfaden hält es ebenso. Beide automatischen Verfahren entscheiden nach einem Informationskriterium, das auf den Daten des Anpassungszeitraums berechnet wird: ETS und ARIMA wählen das einfachste Modell, das kurzfristig noch genaue Prognosen liefert („ETS and ARIMA select the simplest model that is still good at producing accurate short-term forecasts“), und die Autoren ergänzen, ein solches Modell könne weit vom Optimum entfernt sein, wenn der Prognosehorizont deutlich länger als eine Periode ist und die Prognosen mit einem anderen Maß beurteilt werden als dem, mit dem das Modell angepasst wurde („could be far from optimal if the forecasting horizon considered is much greater than one period and the measure used for evaluating the forecasts differs from the one used for training“).

Eine Prognosekombination mittelt die Prognosen mehrerer Verfahren. Der wichtigste Benchmark in M4, Comb, ist das einfache arithmetische Mittel aus SES, Holt und Damped („The simple arithmetic average of SES, Holt and Damped exponential smoothing“). Dazu kamen zwei Benchmarks des maschinellen Lernens, ein Perzeptron (MLP) und ein rekurrentes Netz (RNN), beide von sehr einfacher Architektur und

Parametrisierung („of a very basic architecture and parameterization“) und für jede Reihe einzeln trainiert („trained individually for each series“). Ein hybrides Verfahren führt beide Ansätze zusammen: Das Siegerverfahren in M4 verbindet Formeln der exponentiellen Glättung mit einem rekurrenten neuronalen Netz, das die Prognose erzeugt („mixes ES formulas with a recurrent neural network (RNN) forecasting engine“), und seine Glättungsparameter werden zusammen mit den Gewichten des Netzes und mit demselben Gradientenverfahren geschätzt („are fitted concurrently with the RNN weights by the same gradient descent method“).

Was zeigten 100.000 Reihen über diese Verfahren?

§ 3 von 6

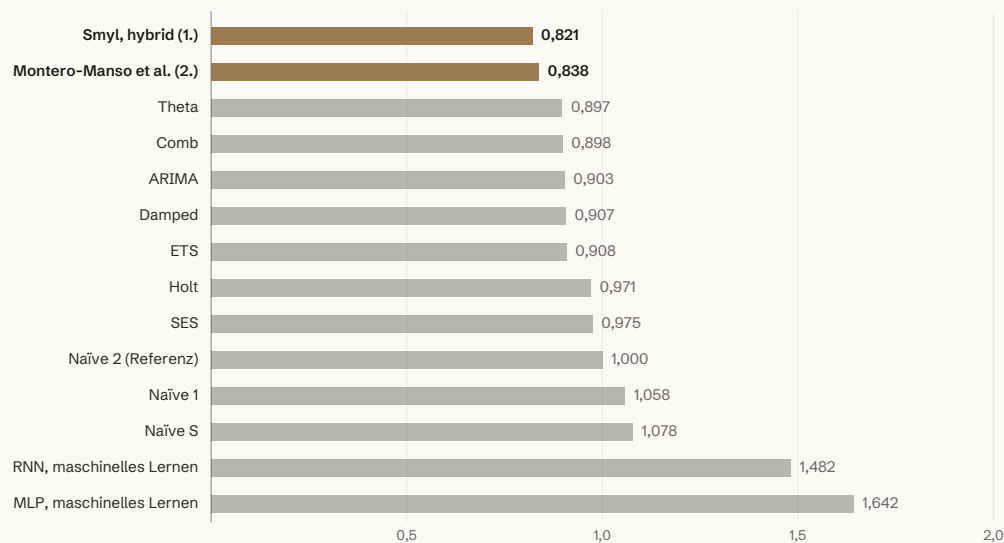
M4 lief von Januar bis Mai 2018 und verlangte Prognosen für 100.000 Reihen: 23.000 jährliche, 24.000 quartalsweise, 48.000 monatliche und 5.000 wöchentliche, tägliche oder stündliche (Tabelle 1, S. 3 der Vorabfassung). Prognostiziert wurden sechs Schritte für Jahres-, acht für Quartals- und 18 für Monatsreihen („six for yearly, eight for quarterly and 18 for monthly series“), dazu 13, 14 und 48 Schritte für Wochen-, Tages- und Stundenreihen; die Veranstalter werteten die Prognosen an den zurückgehaltenen Testdaten aus. Von 248 registrierten Teilnehmenden reichten 49 gültige Prognosen für alle Reihen ein; mit zehn Benchmarks und zwei Vergleichsstandards, ETS und ARIMA, ordnen die Ergebnistabellen 61 Verfahren.

Die Genauigkeit wurde mit zwei Maßen gemessen, die der nächste Abschnitt erklärt: dem symmetrischen mittleren absoluten prozentualen Fehler (sMAPE) und dem mittleren absoluten skalierten Fehler (MASE). Beide wurden erst über die Prognosehorizonte einer Reihe gemittelt und dann über die Reihen, sodass jede Reihe gleich viel zählt („thus weighting all of the series in the dataset equally“). Die Rangfolge beruht auf dem OWA (Overall Weighted Average) der beiden: sMAPE und MASE werden jeweils durch den Wert von Naïve 2 geteilt, und aus den beiden relativen Werten wird das einfache arithmetische Mittel gebildet („We compute the OWA of sMAPE and MASE by first dividing their total value by the corresponding value of Naïve 2 to obtain the relative sMAPE and the relative MASE, respectively, and then computing their simple arithmetic mean.“). Ein OWA von 1 entspricht der Genauigkeit von Naïve 2, kleiner ist besser; im Beispiel der Autoren bedeutet ein OWA von 0,88, dass ein Verfahren unter Berücksichtigung von MASE und sMAPE etwa 12 Prozent genauer ist als Naïve 2 („about 12% more accurate than Naïve 2, taking into account both MASE and sMAPE“). Sie fügen hinzu, sie

wollten nicht behaupten, die gewählten Maße seien die geeignetsten gewesen („We will not pretend that the selected measures were the most appropriate to use“). Abbildung 1 zeigt 14 der 61 Verfahren.

Abbildung 1 Genauigkeit von 14 der 61 M4-Verfahren, relativ zu Naïve 2

Ein hybrides Verfahren und eine Kombination lagen vorn. Theta, Comb, ARIMA, Damped und ETS lagen höchstens 0,011 auseinander; die beiden Benchmarks des maschinellen Lernens waren von den 14 gezeigten Verfahren am ungenauesten.



Waagerechte Achse: OWA, das Mittel aus sMAPE und MASE eines Verfahrens, jeweils geteilt durch den Wert von Naïve 2; Naïve 2 liegt daher definitionsgemäß bei 1, kleiner ist genauer. Comb ist das einfache Mittel aus SES, Holt und Damped, ARIMA und ETS wählen ihre Modelle automatisch, Naïve S ist die saisonale naive Prognose, und RNN und MLP sind die beiden Benchmarks des maschinellen Lernens. Tabelle 4 (S. 8 und 9 der Vorabfassung) nennt kein Intervall für das OWA; zu Damped, Comb und ARIMA schreiben die Autoren, dass zwischen ihnen kaum ein Unterschied statistisch signifikant ist, nach ihrem Test, der die Verfahren nach dem sMAPE ordnet. Die Abbildung lässt die übrigen 47 Einreichungen weg, darunter die beiden ungenauesten aller 61 Verfahren, beide rein statistisch, mit 2,561 und 4,387.

Quelle Makridakis et al. (2020), Tabelle 4, S. 8 und 9 der vorliegenden Vorabfassung (Article in Press, S. 1 bis 21): das OWA (Overall Weighted Average) aus sMAPE und MASE jedes Verfahrens, jeweils geteilt durch den Wert von Naïve 2, über alle 100.000 Reihen. Gezeigt sind die beiden besten Einreichungen, alle zehn Benchmarks und die beiden Vergleichsstandards ETS und ARIMA; die übrigen 47 Einreichungen sind weggelassen. Die Verfahrenstypen sind die, die die Tabelle angibt. Werte der Autoren, keine Neuberechnung.

Das Kombinieren nennen die Autoren ihren wichtigsten Befund: Alle Spitzenverfahren, bei Punktprognosen wie bei Prognoseintervallen, waren Kombinationen überwiegend statistischer Verfahren („all of the top-performing methods, in terms of both PFs and PIs, were

combinations of mostly statistical methods“), und unter den zehn genauesten war nur ein rein statistisches Verfahren („there was only one pure statistical method among the first ten most-accurate ones“), während die zehn ungenauesten durchweg reine Verfahren waren, statistische oder solche des maschinellen Lernens („all of the ten least-accurate methods were either pure statistical or pure ML“); die beiden ungenauesten aller 61, mit 2,561 und 4,387, waren statistische Einreichungen, die Abbildung 1 nicht zeigt (Tabelle 4, S. 9 der Vorabfassung). Dieselbe Tabelle ordnet den Sieger als hybrides Verfahren ein und die nächsten fünf als Kombinationen (S. 8 der Vorabfassung). Comb, das einfache Mittel, kam auf ein OWA von 0,898, gegenüber 0,975 für SES, 0,971 für Holt und 0,907 für Damped, und war damit genauer als jedes der drei Verfahren, aus denen es gebildet ist („Comb provides more accurate PFs than the three individual methods being combined“). Zugleich warnen die Autoren, dass Kombinieren nicht garantiert, Einzelmodelle zu übertreffen („combining does not guarantee that their forecasts will outperform single models“): Theta, ein Einzelverfahren, kam auf 0,897.

Unter den besten Benchmarks und Vergleichsstandards waren die Abstände klein: Theta 0,897, Comb 0,898, ARIMA 0,903, Damped 0,907 und ETS 0,908. Zu Damped, Comb und ARIMA schreiben die Autoren, ihre Genauigkeit sei in M4 sehr ähnlich („are very similar in M4“) und zwischen den Verfahren seien kaum statistisch signifikante Unterschiede zu finden („Fig. 1 indicates that hardly any statistically significant differences among the methods can be found“). Ihr Rückblick auf ARIMA führt zu einem verwandten Punkt: Theoretische Statistiker brauchten lange, um aus den früheren M-Wettbewerben zu akzeptieren, dass die Anpassungsgüte nicht unbedingt mit der Genauigkeit außerhalb des Schätzzeitraums zusammenhängt („goodness of fit was not necessarily related to post-sample accuracy“). Ein Modell, das die Vergangenheit gut abbildet, muss den Testzeitraum also nicht gut prognostizieren.

Das genaueste Verfahren, Smyls Hybrid aus exponentieller Glättung und rekurrentem Netz, kam auf 0,821; seine Verbesserung gegenüber Comb liege nahe an beeindruckenden 10 Prozent, schreiben die Autoren („close to an impressive 10%“), in Tabelle 4 sind es 9,4 Prozent beim sMAPE und 8,6 Prozent beim OWA. Zweiter wurde mit 0,838 eine Kombination aus sieben statistischen Verfahren und einem Verfahren des maschinellen Lernens („a combination of seven statistical methods and a ML one“), deren Gewichte ein Algorithmus des maschinellen Lernens für jede Reihe einzeln bestimmte, trainiert darauf, die Prognosefehler in Tests an zurückgehaltenen Daten zu minimieren

(„calculated by a ML algorithm that was trained to minimize the forecasting errors through holdout tests for each individual series“). Die Verfahren des maschinellen Lernens schnitten schlecht ab: Von den fünf eingereichten Verfahren des maschinellen Lernens, vier reinen und einer Kombination, war keines genauer als der Benchmark Comb und nur eines genauer als Naïve 2 („All of them were less accurate than the Comb benchmark in terms of PFs, and only one was more accurate than the Naïve 2 method.“). Die beiden einfachen Benchmarks des maschinellen Lernens waren mit 1,482 und 1,642 die ungenauesten Verfahren in Abbildung 1. Die Erklärung der Autoren ist eine Vermutung: Die Gründe könnten in der geringen Nutzung von Cross-Learning liegen („could be the limited usage of cross-learning“), also von Informationen aus anderen Reihen, wie sie die drei besten Verfahren nutzten, dazu in kurzen Einzelreihen, Überanpassung und nicht stationären Daten. Sie schränken das Ergebnis selbst ein: Weil es so wenige Einreichungen dieser Art gab, seien womöglich weitere Informationen nötig, bevor sich der Befund verallgemeinern lasse („given the paucity of entries in the ML category, more information might still be needed to make it safe to generalize this finding“). Der Leitfaden zur KI-Nachfrageprognose beschreibt, was ein Prognosesystem um ein solches Modell über seine Genauigkeit hinaus festlegen muss.

M4 bewertete auch Prognoseintervalle, und zwar zum 95-Prozent-Niveau („The M4 Competition adopted a 95% prediction interval (PI) for estimating the uncertainty around the point forecasts.“), für die 20 Einreichungen, die solche Intervalle lieferten, für Naïve 1 und für die beiden Vergleichsstandards ETS und ARIMA. Die beiden besten Verfahren waren auch hier die besten: Ihre Intervalle enthielten 94,78 und 95,96 Prozent der späteren Werte (Tabelle 5, S. 10 der Vorabfassung). Bei den übrigen 21 Verfahren enthielten die Intervalle zwischen 79,68 und 93,39 Prozent der späteren Werte und waren damit zu eng, wie die Intervalle der meisten Verfahren zuvor; die Autoren fordern, die Unterschätzung der Unsicherheit durch diese Verfahren anzugehen („the underestimation of uncertainty by the remaining methods must be addressed“). Punktprognosen allein nennen sie für Entscheidungen unzureichend und oft gefährlich („insufficient and often dangerous to utilize for decision making on their own“).

Warum lassen sich Prognosefehler nicht ohne Weiteres über Reihen mitteln?

§ 4 von 6

Weil jeder Fehler die Einheit seiner Reihe trägt. Hyndman und Koehler (2006)¹ nennen Maße wie den mittleren absoluten Fehler (MAE) und die Wurzel aus dem mittleren quadratischen Fehler skalenabhängig: Sie taugen für den Vergleich verschiedener Verfahren auf denselben Daten, nicht aber für den Vergleich zwischen Datensätzen mit unterschiedlicher Skala („These are useful when comparing different methods applied to the same set of data, but should not be used, for example, when comparing across data sets that have different scales.“). Den MAE über Reihen hinweg zu berechnen, ergebe wegen der unterschiedlichen Skalen keinen Sinn („It makes no sense to compute the MAE across series (due to their different scales).“). Nach der Lesart dieses Leitfadens ist ein Fehler von 50 Einheiten groß für eine Produktlinie, die 100 Stück im Monat verkauft, und klein für eine mit 10.000. Liegt die Prognose für die große Linie um 500 Stück daneben und die für die kleine um 50, beträgt der mittlere absolute Fehler der beiden 275 und wird fast ganz von der großen Linie bestimmt. Der Beitrag zur Forecastgenauigkeit zeigt, warum vorzeichenbehaftete und absolute Fehler, ihre Verteilung, der Horizont und der Nenner getrennt bleiben müssen, wenn die Fehler einer Prognose zusammengefasst werden; hier geht es darum, verschiedene Reihen auf eine Skala zu bringen.

Prozentuale Fehler beseitigen die Einheit und werden häufig verwendet, um die Prognosegüte über Datensätze hinweg zu vergleichen („frequently used to compare forecast performance across different data sets“), teilen aber durch den Istwert. Deshalb werden sie, auch der mittlere absolute prozentuale Fehler (MAPE), unendlich oder undefiniert, wenn ein Istwert null ist, und ihre Verteilung ist extrem schief („extremely skewed“), wenn er nahe null liegt; zudem setzen sie einen sinnvollen Nullpunkt voraus („assume a meaningful zero“). Die symmetrische Variante, der sMAPE, teilt durch Istwert und Prognose zusammen. Das mildert das Problem, beseitigt es aber nicht: Nahe null wird weiterhin durch eine Zahl nahe null geteilt („the measure still involves division by a number close to zero“), der sMAPE kann negativ werden („sMAPE and sMdAPE can take negative values“), und bei gleichem Istwert wird eine zu niedrige Prognose stärker bestraft als eine zu hohe.

Ihre Beispiele machen das greifbar. Eines sind drei Jahre monatlicher Absatz eines Schmierstoffs in Großgebinden („monthly sales of a lubricant product sold in large containers“), eine Reihe mit Monaten

ohne Absatz. Für sie war der MAPE bei allen vier verglichenen Verfahren unendlich oder undefiniert, im Anpassungs- wie im Testzeitraum, und nur der MASE ließ sich für jedes Verfahren berechnen (Tabelle 3, S. 681). Solche Reihen seien nicht ungewöhnlich, betonen sie: Daten sporadischer Nachfrage enthalten oft Nullen, und viele Zeitreihen, die Prognostiker interessieren, enthalten negative Werte („intermittent demand data often contain zeros and many time series of interest to forecasters contain negative observations“). Selbst in einer Reihe ohne Nullen, der M3-Reihe N0472, war ein relatives Maß, das durch die Fehler des naiven Verfahrens teilt, der geometrische Mittelwert der relativen absoluten Fehler, bei jedem Verfahren unendlich oder undefiniert (Tabelle 1, S. 681), weil aufeinanderfolgende Werte gleich waren („the occurrence of consecutive observations which take the same value, something that very often occurs with real data“).

Ihr Vorschlag ist, jeden Fehler daran zu messen, wie gut das naive Verfahren in der Vergangenheit derselben Reihe war. Der skalierte Fehler teilt einen Prognosefehler durch den mittleren absoluten Fehler der naiven Einschrittprognosen im Anpassungszeitraum und ist damit unabhängig von der Skala der Daten („which is clearly independent of the scale of the data“); der MASE ist der Mittelwert der absoluten skalierten Fehler. Liegt er unter eins, waren die Fehler eines Verfahrens im Mittel kleiner als die Einschrittfelder des naiven Verfahrens im Anpassungszeitraum, liegt er über eins, größer. Weil dieser Maßstab eine Einschrittprognose ist, liegen die skalierten Fehler von Prognosen über mehrere Schritte typischerweise über eins („the scaled errors for multi-step forecasts are typically larger than one“); nach der Lesart dieses Leitfadens zeigt ein MASE von 1,5 bei einem Horizont von 18 Monaten daher nicht, dass das Verfahren bei diesem Horizont der naiven Prognose unterlag. Unendlich oder undefiniert wird das Maß nur, wenn alle historischen Werte gleich sind („when all historical observations are equal“). Hyndman und Koehler¹ schlagen skalierte Fehler als Standard für den Vergleich der Genauigkeit über Reihen mit unterschiedlicher Skala vor („the standard approach in comparing forecast accuracy across series on different scales“) und lassen den älteren Maßen Raum: Liegen alle Reihen auf derselben Skala, könne der MAE vorzuziehen sein, weil er leichter zu erklären ist („if all series are on the same scale, then the MAE may be preferred because it is simpler to explain“), und sind alle Daten positiv und weit größer als null, könne der MAPE der Einfachheit halber vorzuziehen sein („If all data are positive and much greater than zero, the MAPE may still be preferred for reasons of simplicity.“).

M4 verwendete eine Variante. Ihr MASE skaliert mit der saisonalen naiven Prognose, mit einer Saisonlänge von 12 für Monats-, vier für Quartals-, 24 für Stunden- und eins für Jahres-, Wochen- und Tagesdaten („12 for monthly, four for quarterly, 24 for hourly and one for yearly, weekly and daily data“); ursprünglich, so die Autoren, war m unabhängig von der Frequenz der Daten auf eins gesetzt („originally m was set equal to one regardless of the frequency of the data“). Probleme mit dem sMAPE vermied M4 schon durch seine Anlage: Die Reihen wurden so angepasst, dass keine negativen Werte und keine Werte unter 10 vorkamen („scaled to prevent negative observations and values lower than 10“), durch Addition einer Konstante, wo nötig, was 29-mal geschah, und Reihen mit geringem Volumen oder sporadischer Nachfrage wurden nicht aufgenommen („low-volume and intermittent time series were not considered for the M4 dataset“). Für die eigenen Produktlinien eines Unternehmens gilt diese Garantie nicht.

Was sollte ein Prognosevergleich festlegen, bevor der erste Wert berechnet wird?

§ 5 von 6

Sechs Dinge, festgelegt, bevor ein Verfahren bewertet wird: die Reihe mit Einheit und Frequenz; den Testzeitraum, den kein Verfahren sehen darf; die Horizonte, gewählt nach der Entscheidung, der die Prognose dient; die Benchmarks; das Fehlermaß und seine Skalierung; und wie das Ergebnis berichtet wird, im Vergleich mit jedem Benchmark bei jedem Horizont. Die Felder folgen der Anlage von M4, die Testdaten zurückhielt, ihre Horizonte nach der Art der Entscheidungen festlegte, die eine Datenfrequenz am ehesten stützt („based on the nature of the decisions that each frequency of data is most likely to support“), und jede Einreichung mit Verfahren bekannter Eigenschaften verglich („of well-known properties“), sowie dem Plädoyer von Hyndman und Koehler¹ für skalierte Fehler. Der Beitrag zu Forecast Value Added wendet dieselbe Disziplin auf eine Änderung an, die Menschen an einer statistischen Basisprognose vornehmen.

Abbildung 2 Ein Blatt für einen Prognosevergleich

Reihe, Testzeitraum, Horizont, Benchmarks, Fehlermaß und die Form des Ergebnisses festhalten, bevor ein Verfahren bewertet wird.

Reihe	Testzeitraum	Horizont	Benchmarks	Fehlermaß	Ergebnis
Welche Reihe, Einheit, Frequenz?	Welche letzten Werte bleiben ungesehen?	Wie viele Schritte, und warum?	Welche naiven Verfahren, welches Mittel?	Skaliert? Werte bei oder nahe null?	Besser als jeder Benchmark, je Horizont?

Die Werte der Benchmarks neben den Wert des Verfahrens stellen, Horizont für Horizont.

Eine Zeile je Reihe oder Gruppe von Reihen. Testzeitraum und Benchmarks festlegen, bevor das erste Verfahren angepasst wird. Ein Wert ohne die Werte seiner Benchmarks lässt sich nicht einordnen.

Quelle Entscheidungsblatt des Autors. Die Felder für Testzeitraum und Horizont folgen Makridakis et al. (2020), S. 4 und 5 der vorliegenden Vorabfassung (Article in Press); das Feld für die Benchmarks folgt Makridakis et al. (2020), S. 3 und 4 dieser Fassung; das Feld für das Fehlermaß folgt Hyndman und Koehler (2006), S. 682 bis 686, und Makridakis et al. (2020), S. 5 dieser Fassung. Die Felder für Reihe und Ergebnis und die Regel in der Fußzeile stammen von diesem Leitfaden. Die Felder sind Eingaben der Lesenden; es werden keine Unternehmensdaten geliefert.

Ein Beispiel ohne Unternehmensdaten: Ein Hersteller prognostiziert den monatlichen Bedarf in Stück für 40 Ersatzteillinien, von wenigen Stück bis zu mehreren Tausend im Monat, mit drei Jahren Historie. Das Blatt nennt für jede Linie Einheit und Frequenz, hält die letzten zwölf Monate als Testzeitraum zurück, mit einem einzigen Prognosezeitpunkt je Linie wie in M4, und leitet die Horizonte aus den Entscheidungen ab: drei Monate für die Produktionsplanung, zwölf für das Budget. Jedes infrage kommende Verfahren wird mit der saisonalen naiven Prognose, mit Naïve 2 und mit einem einfachen Mittel aus Verfahren der exponentiellen Glättung verglichen, weil dieses Mittel in M4 genauer war als jedes seiner drei Bestandteile. Die Fehler werden je Linie nach Art des MASE skaliert, weil ein über alle Linien gepoolter mittlerer absoluter Fehler von den größten Linien bestimmt würde und weil einige kleine Linien Monate ohne Bedarf

haben, in denen ein prozentualer Fehler undefiniert ist. Für diese Linien hält das Blatt auch fest, dass M4 solche Reihen ausließ und keine der beiden Quellen Verfahren für sporadische Nachfrage prüft. Das sind Beispiele des Autors, sie sind nicht geprüft.

Welche dieser Ergebnisse hielten einer kurzen oder sporadischen Reihe nicht stand?

§ 6 von 6

Die Quellen beantworten eine engere Frage, als sie sich in der Umsatz- oder Bedarfsplanung stellt. M4 ist ein einzelner, einmal durchgeführter Wettbewerb mit kontinuierlichen Reihen, und die Autoren sagen das selbst: Ihre Befunde beziehen sich auf kontinuierliche betriebliche Reihen, einige gelten womöglich nicht für Reihen mit geringem Volumen oder sporadischer Nachfrage („the findings of M4 refer to continuous business series, meaning that some of them may not apply for low-volume or intermittent series“). Die Reihen kamen ohne Startdatum, Reihen mit Werten unter 10 wurden angehoben, und jede wurde von einem einzigen Zeitpunkt aus prognostiziert, mit Verfahren, die keine andere Variable sahen. Das Teilnehmerfeld wählte sich selbst aus, denn viele merkten, dass ihre Verfahren nicht genauer waren als die Benchmarks, und reichten deshalb nicht ein („Many participants realized that their methods were no more accurate than the benchmarks, and therefore decided not to submit their forecasts“), und der Befund zum maschinellen Lernen beruht auf fünf Einreichungen aus dem Jahr 2018, was nach den Autoren für eine Verallgemeinerung womöglich zu wenig ist. Spätere Prognosewettbewerbe gehören nicht zu den Quellen dieses Leitfadens. Hyndman und Koehler (2006)¹ argumentieren aus den Definitionen der Maße und führen die Probleme an drei Beispielreihen und den M3-Daten vor; welches Maß zu besseren Entscheidungen führt, prüfen sie nicht.

Nach der Lesart dieses Leitfadens lassen sich drei Dinge mit einiger Zuversicht übertragen: jedes Verfahren mit benannten naiven Benchmarks und einer einfachen Kombination vergleichen, bevor man ihm traut; Verfahren an einem Testzeitraum beurteilen, den sie nicht gesehen haben, nicht an der Anpassung an die Vergangenheit; und die Genauigkeit über Reihen hinweg mit einem skalierten Fehler vergleichen. Weniger sicher ist die Rangfolge der Verfahrensfamilien für die Reihen eines bestimmten Unternehmens. In M4 lagen die besten Benchmarks und Vergleichsstandards dicht beieinander, und zwischen Damped, Comb und ARIMA war nach dem Test der Autoren kaum ein Unterschied statistisch signifikant; die Sieger kombinierten Verfahren oder verbanden statistische Teile mit solchen des maschinellen Lernens, und am ungenauesten waren reine Verfahren beider Arten.

Wie sich all das bei kurzen Umsatzreihen verhält, bei Produkten mit Monaten ohne Nachfrage oder bei einer Prognose, die auch eine Pipeline oder eine Preisänderung nutzen sollte, sagen die Quellen nicht.

 Ende des Beitrags

Referenzen

Zitate stehen in der Sprache des Originals; die deutschen Wiedergaben im Fließtext sind sinngemäß und stammen vom Autor. Seitenangaben zu Makridakis et al. (2020) beziehen sich auf die vorliegende Vorabfassung (Article in Press) mit den Seiten 1 bis 21; Jahrgang, Heft und Seiten unten stammen aus dem Crossref-Datensatz.

- 1 Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688.
<https://doi.org/10.1016/j.ijforecast.2006.03.001>
- 2 Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1), 54–74.
<https://doi.org/10.1016/j.ijforecast.2019.04.014>

Zitierweise

Isoglu, S. (2026, 18. Oktober). Was ist Zeitreihenprognose? Sinan Isoglu.
<https://isoglu.com/de/insights/journal/was-ist-zeitreihenprognose/>
Online-Fassung: <https://isoglu.com/de/insights/journal/was-ist-zeitreihenprognose/>

Sinan Isoglu, MBA (Quantic)

Führungskraft für Umsatzwachstum, Dozent und Doktorand
© 2026 Sinan Isoglu · isoglu.com · Alle Rechte vorbehalten