Bevölkerung

Sterbefallzahlen schätzen für Berlin – ein Werkstattbericht

von **Kerstin Erfurth** und **Jörg Höhne**

Verzögerte Meldungen der Standesämter hinsichtlich der Sterbefälle können die Bereitstellung aktueller und belastbarer Daten durch die Statistischen Ämter des Bundes und der Länder beeinträchtigen. Dies führt dazu, dass die Fälle für einen gewissen Zeitraum nicht vollständig zur Verfügung stehen und auf diese Weise keine Aussage über die aktuelle Situation im Sterbegeschehen liefern. Es wurde ein Modell gesucht, welches in der Lage ist, aus verfügbaren Daten Wissen zu extrahieren und dieses für eine Schätzung zu verwenden. Aus diesem Wissen kann eine Prognose für die jüngste Vergangenheit erstellt werden. Derartige Prognosen, welche sich auf derzeitige statt zukünftige Zeitpunkte beziehen, werden als Nowcast bezeichnet. Sie können einen besseren Hinweis zur aktuellen Situation geben als rohe, unfertige Fallauszählungen. Auf der Suche nach einem Modell hat sich eine gemischte Methode zur Vervollständigung der fragmentarischen Sterbefallmeldungen herauskristallisiert, welche Elemente des maschinellen Lernens verwendet.

Einführung

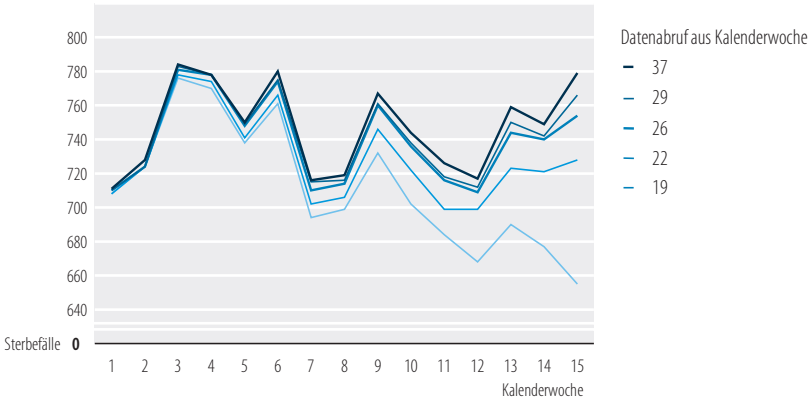
Aktuelle Meldungen zu Sterbefallzahlen interessieren uns – gerade jetzt mitten in einer Pandemie. Jeder schaut aufmerksam auf die Entstehung und den Verlauf der Zahlen, die ein wichtiger Indikator bei der Beobachtung des aktuellen Geschehens sind. Doch nicht nur die Zahlen derjenigen Personen, die direkt an COVID-19 gestorben sind, treiben uns um. Es sind die Sterbefallzahlen insgesamt, die viel verraten: unter anderem, wie sich Veränderungen und Maßnahmen im Zusammenhang mit der Corona-Pandemie auf unsere Gesundheit und demzufolge auch auf unsere Sterblichkeit auswirken.

Doch wie verlässlich sind diese aktuellen, derart stark nachgefragten Sterbefallzahlen? Leider herrscht insbesondere für das Sterbegeschehen in Berlin eine große Unsicherheit bei den aktuellen Fallzahlen. Das Wissen um die Verstorbenen ist extrem unvollständig. Es vergeht eine gewisse Zeit, bis ein Sterbefall beim Standesamt gemeldet und diese Meldung dort so bearbeitet ist, dass sie in der amtlichen Statistik registriert wird. Insbesondere in Berlin verstreichen teilweise mehrere Wochen bis zur vollständigen Erfassung aller Sterbefälle.

Abbildung a vermittelt einen Eindruck davon, wie stark sich Sterbefallmeldungen aus verschiedenen Datenständen unterscheiden können. Zu sehen sind die eingegangenen Sterbefallmeldungen für die Kalenderwochen 1 bis 15 des Jahres 2020 zu verschiedenen Zeitpunkten des Datenabrufs. Dabei ist der Datenstand vom 8. Mai 2020 (Kalenderwoche 19) der früheste, der veröffentlicht wird. Der jüngste hier dargestellte Datenstand ist vom 11. September 2020, also aus Kalenderwoche 37. Zwischen beiden Datenbeständen zeigen sich starke Abweichungen.[1]

Diese Unsicherheit wurde zum Anlass genommen, nach einer Möglichkeit zu suchen, die gegenwärtigen Sterbefallzahlen abzuschätzen. Auch wenn

a | Eingegangene Sterbefallmeldungen (Fallauszählungen) 2020 in Berlin nach Kalenderwochen



1 Die letzten vier Wochen, also die Kalenderwochen 16 bis 19, werden nicht dargestellt, da für diese bereits davon ausgegangen wird, dass die Fallzahlen unvollständig und dementsprechend unsicher sind. Diese Unvollständigkeit zeigt sich jedoch auch weit über die letzten vier Wochen hinaus.

geschätzte Zahlen keineswegs korrekte und geprüfte Ergebnisse ersetzen können, so können sie zumindest in dem Zeitraum unterstützen, in dem die Wahrheit noch nicht bekannt ist. Ähnlich wie bei der Wettervorhersage gibt uns eine Schätzung ein Gefühl von Sicherheit, Planbarkeit und Kontrolle.

Für den hier vorgestellten Ansatz zur Schätzung von Sterbefallzahlen am aktuellen Rand in Berlin werden Elemente des maschinellen Lernens (ML) genutzt, wenngleich dieser Ansatz nicht als reines ML-Verfahren interpretiert werden kann. Generell wird für Verfahren im Bereich des maschinellen Lernens eine Ausgangssituation mithilfe von Input-Werten beschrieben. In diesem Fall sind dies die bisher eingegangenen (historischen) Sterbefallmeldungen. Begibt man sich darüber hinaus in die Unterkategorie des *Supervised Learnings*, also des überwachten Lernens, werden zusätzlich auch Output-Werte herangezogen. Sie beschreiben „wahre" Ergebnisse, welche beim Trainieren und Testen des Modells als Zielgröße genutzt werden. In diesem Anwendungsfall entspricht die Zielgröße der tatsächlichen Anzahl an Sterbefällen in Berlin für bestimmte Kalenderwochen.

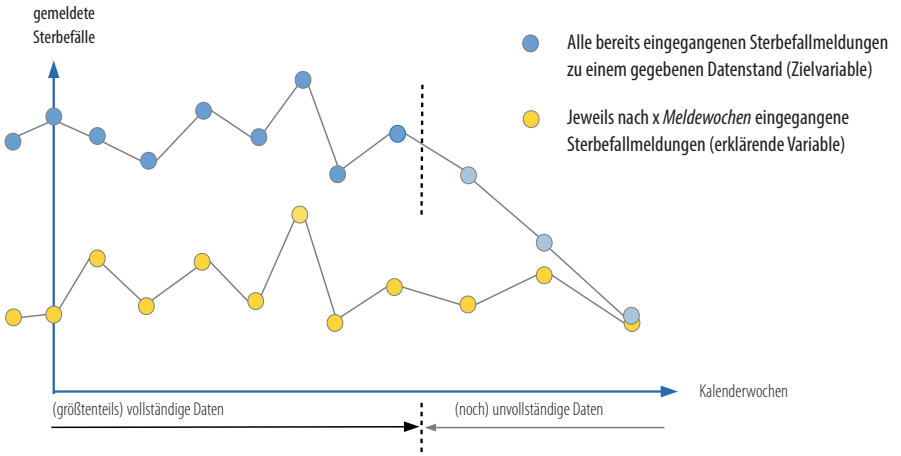
Datengrundlage

Als (Trainings-)Datensatz wurden die Sterbefallmeldungen beginnend am 1. Januar 2017 (Kalenderwoche 1) bis zum 23. Juni 2020 (Kalenderwoche 26) verwendet. Dabei sind nur die Merkmale Sterbedatum, Meldedatum und die Nummer des Standesamtes, das den Sterbefall erfasst hat (Sterbebezirk), relevant. Aus dieser Datenbasis lassen sich Sterbefallzahlen für jede Kalenderwoche und jeden Bezirk aggregieren. Die Input-Werte des Modells sind demzufolge Fallzahlen, welche nach einem bestimmten Zeitraum bereits erfasst respektive eingegangen sind (zum Beispiel nach einer Woche, nach zwei Wochen, nach drei und so weiter). Nach einer hinreichend langen Zeit entsprechen diese Fallzahlen den zu schätzenden Sterbefallzahlen (Output-Werte) und werden als „wahre Werte" festgelegt.

Nehmen wir beispielsweise das Sterbedatum 1. Januar 2020. Da einige Schritte vom Tod einer Person bis zur Meldung erforderlich sind, werden, bis auf wenige Ausnahmen, an diesem Tag selbst keine Sterbemeldungen für eben diesen Tag eingegangen sein. Etwas anders sieht es beispielsweise am 8. Januar 2020 aus: Eine Woche später liegen bereits mehr Sterbemeldungen vom 1. Januar 2020 vor. Am 15. Januar 2020 liegen dann noch einmal mehr Meldungen vor und so weiter. Dieses Prinzip lässt sich auch auf Kalenderwochen anwenden. In der zweiten Kalenderwoche können die Fälle ermittelt werden, welche in der ersten verstorben sind. Der Vorteil der Betrachtung von Sterbewochen statt Sterbetagen besteht darin, dass dadurch der Unterschied zwischen den Wochentagen (insbesondere den Wochenenden) eliminiert wird. Dieser „Rückblick" wird im Folgenden mit „nach x Wochen eingegangene Meldungen" bezeichnet. Die Anzahl der nach x Wochen eingegangenen Sterbefallmeldungen kann entsprechend für alle vergangenen Kalenderwochen bestimmt werden. So lässt sich beispielsweise berechnen, wie viele Meldungen für die Sterbewoche 46 im Jahr 2018 (12.–18. November 2018) nach zwei Wochen bereits gemeldet beziehungsweise bearbeitet wurden, obwohl die Fallzahlen aus heutiger Perspektive bereits vollständig sind. Wie weit in die Vergangenheit geschaut werden kann, ist für die Schätzung von Relevanz, da mit einem längeren Blick in die Vergangenheit auch mehr Daten zur Verfügung stehen. Dieses Mehr an Daten respektive Informationen sollte selbstverständlich bei der Modellierung berücksichtigt werden. Andersherum können jedoch nur die Informationen verwendet werden, die auch vorliegen. Wird beispielsweise die gerade vergangene Kalenderwoche als Sterbewoche betrachtet, so sind die nach zwei Wochen eingegangenen Sterbefallmeldungen noch nicht verfügbar.

Eine schematische Darstellung der Datengrundlage zeigt Abbildung b. Zu sehen ist, wie der Datensatz in zwei Teile aufgeteilt werden kann: Zum einen die vollständigen Daten, die bis zu einem gewissen Zeitpunkt als „wahr" angenommen werden (blaue Punkte). Sie entsprechen dem Datenstand zu einer bestimmten Kalenderwoche, analog Abbildung a. Es handelt sich hierbei um alle bis zur betrachteten Kalenderwoche eingegangenen Sterbefallmeldungen. Wie bereits erwähnt, sind diese Daten ab einem gewissen Zeitpunkt unvollständig. Zum anderen kann

b | Schematische Darstellung der Datengrundlage



der Datenstand aller Sterbemeldungen (blaue Punkte) so gefiltert werden, dass eine weitere Zeitreihe entsteht (gelbe Punkte). Sie stellt eine Art Zwischenstand der eingegangenen Daten dar. Die Filterung wird so durchgeführt, dass für jede Kalenderwoche nur Meldungen berücksichtigt werden, die für eine gegebene Anzahl x an Wochen (*Meldewochen*), beispielsweise nach zwei Wochen, jeweils eingegangen waren. Die Anzahl x ist fest und für jede Kalenderwoche gleich. Beispielsweise beschreibt der erste orange Punkt die nach zwei Wochen eingegangenen Sterbefallmeldungen aus Kalenderwoche 1. Der zweite orange Punkt wiederum die nach zwei Wochen eingegangenen Sterbefallmeldungen aus Kalenderwoche 2 und so weiter.

Lineare Regression

Die nach x Wochen eingegangenen Meldungen sind neben den Angaben zum Bezirk und der Kalenderwoche die einzigen Input-Informationen, die für eine Schätzung der tatsächlichen Sterbefallzahl (Output) herangezogen werden können. Daher ist die Idee, genau diese Sterbefälle, welche dem Statistischen Amt bis zu einem festgelegten Zeitpunkt für eine Kalenderwoche *i* gemeldet wurden, auf die Anzahl der tatsächlich in dieser Kalenderwoche verstorbenen Personen zu regressieren. Die unvollständige Fallzahl z_i beschreibt entsprechend die erklärende Variable und die, nach hinreichend langer Zeit beobachtete, tatsächliche Sterbefallzahl y_i stellt die Zielvariable dar. Der Fehler ε_i gibt den Abstand zwischen wahrem Wert und der Regressionsgeraden an. Bei einer Schätzung $\hat{y}_i$ kann der Fehler nicht angegeben werden.

$$y_i = az_i + b + \varepsilon_i$$

Das bedeutet, dass der tatsächliche Wert der Sterbefallzahlen auf Basis eines Gewichts beziehungsweise Anstiegs a für die bisher eingegangenen Fallzahlen und einem Absolutwert b geschätzt werden kann. Dies kann wie folgt interpretiert werden: Werden beispielsweise die bisher $z_i = 100$ eingegangenen Fallmeldungen mit dem Faktor $a = 6$ multipliziert und anschließend $b = 80$ Fälle dazu addiert, dann entsteht ein Schätzwert von $\hat{y}_i = 680$ Sterbefällen, welche tatsächlich in einer Kalenderwoche *i* eingetreten sein könnten, jedoch noch nicht vollständig registriert wurden. Das Gewicht a gibt mit 1/a quasi einen anteiligen „Bearbeitungsstand" und der Achsenabschnitt b eine Konstante an.

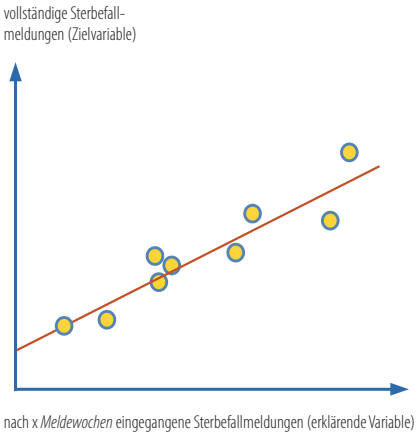
Abbildung c zeigt, wie die vollständigen Sterbefallmeldungen zu einem gegebenen Datenstand (Zielvariable) gegen einen Zwischenstand der Daten – die jeweils nach x Wochen eingegangenen Sterbefallmeldungen (erklärende Variable) – abgetragen werden können. Im Vergleich zu Abbildung b sind die Datenpunkte nun nicht mehr nach Kalenderwochen sortiert, wenngleich jeder einzelne Punkt für eine bestimmte Kalenderwoche steht. Die rote Linie stellt die Regressionsgerade dar und damit den Zusammenhang zwischen den tatsächlichen Sterbefallmeldungen und dem definierten Zwischenstand der Daten.

Parameter der linearen Regression

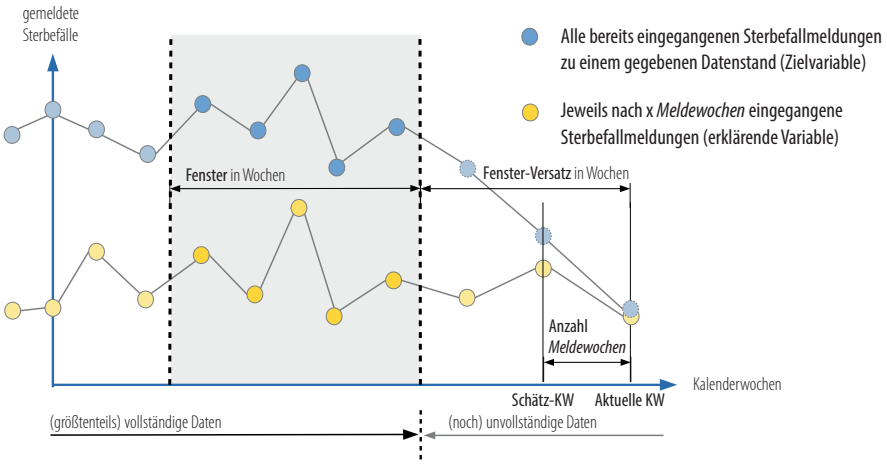
Die beschriebene Modellierung der Regressionsgeraden passiert auf einem bereits in der Vergangenheit liegenden Zeitraum, für welchen die wahren Werte für *y* bekannt sind. Dieser Zeitraum liefert somit einen Trainingsdatensatz, auf dem das lineare Modell generiert wird. Es werden allerdings nur Teile des gesamten Datensatzes, beziehungsweise der Zeitreihe, in die jeweilige lineare Regression eingesetzt. Welcher Teil beziehungsweise welcher Zeitraum in die Berechnungen eingeht, wird durch drei Parameter bestimmt: das *Fenster*, den *Fenster-Versatz* und die Anzahl der Wochen, welche Daten für die aktuelle Kalenderwoche liefern konnten, im Folgenden auch als „Anzahl *Meldewochen*" bezeichnet. Abbildung d illustriert diese Parameter. Darüber hinaus hängt der Zeitraum auch von der aktuellen Kalenderwoche ab; beziehungsweise von der Kalenderwoche, für die ein Schätzwert erstellt werden soll.

Die drei Parameter, die das betrachtete Zeitfenster definieren, werden im Folgenden erläutert. Der erste wird durch die eigentliche Länge des Zeitfensters in Wochen beschrieben. Er soll im Folgenden kurz als *Fenster* bezeichnet werden. Dieses *Fenster* gibt die Länge des Zeitraums an, welcher für die Berechnung der linearen Regression einbezogen wird und

c | Schematische Darstellung der Regressionsgeraden



d | Schematische Darstellung der Parameter des Zeitintervalls für die Regressionsschätzung



enthält die eigentlichen Trainingsdaten. Ein großes *Fenster* nutzt entsprechend mehr Datenpunkte, welche die Regressionsschätzung stabiler machen. Andererseits werden mit einem großen Zeitfenster auch ältere, weniger aktuelle Daten in das Modell einbezogen. Daten in sehr ferner Vergangenheit besitzen, durch eventuell geändertes Meldeverhalten, für den aktuellen Rand möglicherweise keine Vorhersagekraft mehr.

Der zweite Parameter beschreibt den Abstand des *Fensters* zur aktuellen Kalenderwoche und wird als *Fenster-Versatz* (in Wochen) bezeichnet. Dieser zeitliche Versatz stellt genau die Anzahl an Wochen am aktuellen Rand dar, welche bewusst *nicht* in die Modellierung einbezogen werden sollen, für den Fall, dass bei der Schätzung die Sterbefallzahlen für diese Wochen noch nicht vollständig vorliegen. Die unvollständigen Werte würden das Modell folglich „falsch trainieren". Je größer der *Fenster-Versatz,* desto höher ist die Wahrscheinlichkeit, dass die betrachteten Daten bereits endgültig sind. Dem gegenüber steht jedoch die Wahl eines kleinen Versatzes, der den Vorteil mit sich bringt, dass aktuelle Änderungen des Meldeverhaltens im Modell besser berücksichtigt werden können. Es besteht auch hier ein „Tradeoff" zwischen der Unvollständigkeit und der Aktualität der betrachteten Daten. Es ist also entscheidend, das Optimum beider Extrema zu finden – der Kern der dargestellten Analyse.

Der letzte Parameter, welcher die Schätzung beeinflusst, ist die Anzahl der Wochen, welche Daten für die aktuell betrachtete Woche liefern. Dieser Parameter soll mit Anzahl *Meldewochen* bezeichnet werden. Diese Anzahl x an *Meldewochen* definiert, wie die Daten gefiltert werden (gelbe Punkte). Zusätzlich definiert sie, für welche Woche, rückblickend von der aktuellen Kalenderwoche, eine Schätzung frühestens erstellt werden kann – die sogenannte Schätz-Kalenderwoche. Wird eine Schätzung für drei Wochen, also drei *Meldewochen* zurück, vorgenommen, so können auch nur maximal Daten von drei Wochen einbezogen werden. Mehr Daten sind zu diesem Zeitpunkt noch nicht verfügbar.

Für jede geschätzte Woche am aktuellen Rand können jeweils unterschiedliche Parameterkonstellationen entstehen. Die Schätzung für „vor einer (die letzte) Woche" entsteht beispielsweise durch eine lineare Regression, welche als *Fenster* eine Länge von acht Wochen, als *Fenster-Versatz* vier Wochen und verabredungsgemäß eine *Meldewoche* besitzt. Im Gegensatz dazu kann die Schätzung für die vorletzte Woche, also zwei *Meldewochen*, völlig andere Parameter besitzen: beispielsweise *Fenster* = 20 Wochen und *Fenster-Versatz* = 6 Wochen. An dieser Stelle sei angemerkt, dass der *Fenster-Versatz* nie kleiner als die Anzahl der *Meldewochen* sein darf, da er die unvollständigen Meldungen ausschließen soll. Die Unterschiede in der Parameterwahl zwischen den einzelnen Schätzergebnissen kommen dadurch zustande, dass letztendlich jeweils der – historisch betrachtet – beste Parametersatz als Ergebnis des Modelltrainings verwendet wird.

Nun könnte eine lineare Regression auch auf den gesamten Datensatz, also über den kompletten Zeitraum, angewendet und die Koeffizienten a und b einmalig bestimmt werden. Davon wurde allerdings abgesehen, da dieses Vorgehen einen entscheidenden Nachteil besitzt: Änderungen des Meldeverhaltens über die Zeit können auf diese Weise nicht berücksichtigt werden. Angenommen, dass mit den Sommermonaten und der Urlaubszeit weniger Personal für die Meldung und Bearbeitung von Sterbefällen verfügbar ist und den Statistischen Ämtern der Länder dementsprechend prozentual weniger Sterbefälle gemeldet werden. Eine Regression über alle Werte würde dann ein gemitteltes Meldeverhalten verwenden, welches sich im Durchschnitt immer gleich verhält. Ein Blick in die Daten zeigt jedoch ein anderes Bild: Das Meldeverhalten schwankt über die Zeit. Deshalb sollen nur zeitnahe Information Verwendung finden und ein fest definiertes zurückliegendes Zeitfenster für die Berechnung der Regressionsgeraden genutzt werden.

An dieser Stelle soll das Verfahren noch einmal gezielt in seine zwei Bestandteile aufgeteilt betrachtet werden: Der erste Teil (die Analyse) ist die Erstellung von Schätzungen in der Vergangenheit, um die besten Parameter des Modells (*Fenster*, *Fenster-Versatz*, *Meldewochen*) zu finden und die Qualität des Verfahrens zu beurteilen. Der zweite Teil beschreibt die bloße Anwendung des Verfahrens mit den in der Analyse gefundenen Parametern. Bei der reinen Anwendung des Verfahrens ist die Kalenderwoche, für die eine Schätzung erstellt werden soll, in aller Regel der aktuelle Rand. Bei der vorherigen Parametersuche werden im Rahmen des Trainings auch Schätzungen für Kalenderwochen erstellt, die bereits viel weiter in der Vergangenheit liegen. Zur Gütebeurteilung werden diese Schätzungen dann mit den tatsächlich eingetretenen Sterbefällen verglichen. In diesem Teil der Untersuchung gleitet das Fenster entsprechend über die gesamte Zeitreihe. Dabei ist zu beachten, dass der Parameter *Meldewochen* im Anwendungsteil des Verfahrens nicht mehr modifizierbar ist und somit streng genommen keinen veränderlichen Parameter mehr darstellt. Er ist nur während der Analyse variabel. In der praktischen Anwendung ist die jeweilige Anzahl an *Meldewochen* direkt durch die aktuelle Kalenderwoche vorgegeben und muss nicht weiter gesteuert oder verändert werden.

Bis zu diesem Punkt wurde der grobe Ansatz zur Erstellung von Nowcast-Schätzwerten mithilfe einer linearen Regression und ihren Besonderheiten beschrieben. Doch stellt sich die Frage, warum diese Herangehensweise überhaupt funktionieren kann? Das angewendete Prinzip funktioniert nur dann gut, wenn die bisher eingegangenen Sterbefallmeldungen auch möglichst gut mit den (später festgestellten) wahren Fallzahlen korrelieren. Das bedeutet, dass der Verlauf der Zeitreihe der Sterbefallzahlen ähnlich dem Verlauf der Zeitreihe der Sterbefälle sein sollte, die beispielsweise nach jeweils zwei *Meldewochen* eingegangen sind (siehe Abbildung b). Dies kann mit einer linearen Regression erfasst werden. Je stärker der lineare Zusammenhang ausgeprägt ist, umso verlässlicher ist auch die Schätzung für zukünftige Werte. Ist die Korrelation dagegen sehr gering, der lineare Zusammenhang also nicht ausgeprägt,

beziehungsweise liegen die Daten nicht auf einer Regressionsgeraden (siehe Abbildung c), so ist der daraus abgeleitete Schätzwert unsicher und besitzt entsprechend weniger Vorhersagekraft.

Die Anpassungsgüte einer linearen Regression wird mithilfe des Bestimmtheitsmaßes R^2 gemessen. Die Merkmale entsprechen der erklärenden Variable *z* und der Zielvariable *y*. Dabei gibt ein R^2 von 0 an, dass kein linearer Zusammenhang existiert. Ein R^2 von 1 bedeutet, dass ein perfekter linearer Zusammenhang zwischen den beiden Merkmalen (hier die Sterbefallzahlen zu verschiedenen Zeitpunkten der Erfassung) besteht. Die Datenpunkte der beiden Variablen liegen dann exakt auf einer Geraden.

Das Funktionieren der Methode setzt also eine gut ausgeprägte Korrelation zwischen erklärender Variable und Zielvariable voraus. Was kann allerdings unternommen werden, wenn keine oder nur eine sehr schlechte Korrelation zwischen den beiden Zeitreihen erkennbar ist? So ist es beispielsweise möglich, dass die temporären, summierten Berliner Fallzahlen nach zwei *Meldewochen* nur einen schlechten linearen Zusammenhang mit den wahren Fallzahlen für ganz Berlin aufzeigen. Um das Verfahren robuster gegenüber gering korrelierten Zeitreihen zu machen, kann eine weitere Information genutzt werden: die Bezirke, in denen die Sterbefälle registriert wurden.

Kombination aller Bezirke

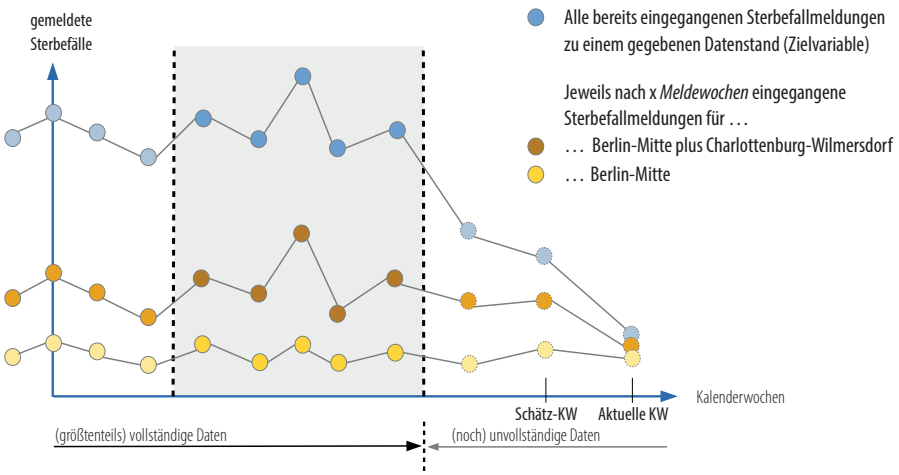
Bei der Verwendung der Sterbefallzahlen auf Bezirksebene ist der Grundgedanke, dass die bereits gemeldeten Fallzahlen aus einem einzelnen Bezirk möglicherweise besser mit der wahren Zahl aller Verstorbenen in Berlin korrelieren als die bisher eingegangene Gesamtzahl aller Bezirke. Denkbar ist auch, dass die Summe der bisher gemeldeten Sterbefälle zweier Bezirke, beispielsweise Berlin-Mitte und Charlottenburg-Wilmersdorf, eine bessere Vorhersage für ganz Berlin liefert. Es können auch weitere Bezirke als Teilmenge Berlins betrachtet werden. De facto kann jede Kombination der zwölf Berliner Bezirke dahingehend untersucht werden, ob ihre Summe besser mit den wahren Werten zusammenhängt. Dieser Aspekt wurde entsprechend in die Untersuchung integriert: Das Verfahren berechnet systematisch alle Kombinationen der zwölf Bezirke und summiert sie jeweils auf. Diese Summen gehen nun als erklärende Variable in die jeweiligen Regressionsberechnungen ein (Abbildung e).

Doch warum kann es passieren, dass weniger Daten bessere Ergebnisse liefern können? Hier spielen mehrere Faktoren eine Rolle. Der wesentliche Aspekt ist dabei das unbeständige Meldeverhalten. So können starke Schwankungen des Meldeverhaltens innerhalb eines Bezirks die Gesamtzahl der bisher eingegangenen Sterbefallmeldungen derart „stören", dass nur eine weniger robuste Aussage für ganz Berlin getroffen werden kann – die Korrelation sinkt entsprechend. Werden solche Störfaktoren („Stör-Bezirke") ausgeschlossen, kann trotz der geringeren einbezogenen Fallzahl ein stabileres Verhältnis entstehen, welches sich in der Regressionsschätzung in einem höheren R^2 widerspiegelt. Dies hat auch eine bessere Qualität der Schätzergebnisse zur Folge.

Die Berücksichtigung dieser Bezirkskomponente lässt das Verfahren abstrakter werden. Es geht nun nicht mehr primär darum, auf Basis eines begründbaren kausalen Zusammenhangs eine Vermutung aufzustellen, mit welchem Bezirk die Zahlen besser geschätzt werden können und warum. Es werden kurzerhand alle Bezirkskombinationen systematisch durchprobiert. Die beste Kombination hinsichtlich des höchsten R^2 wird als Gewinner deklariert und für die weitere Berechnung gewählt. Das Vorgehen lässt sich entsprechend weniger als plausible, „beweisbare" Methodik interpretieren. Für die Anwendung stellt dies jedoch kein Problem dar. Verfahren des maschinellen Lernens lösen sich im Allgemeinen von dem Gedanken, beweisbare Modelle zu erzeugen. Basierend auf vorgegebenen Regeln generieren sie aus vorhandenen (Trainings-)Daten ein abstraktes Modell, wobei kausale Interpretationen keine Rolle spielen. Was zählt, ist die Qualität der Ergebnisse.

Zusammenfassend lässt sich das Schätzverfahren als Kombination einzelner Ansätze beschreiben. Es setzt sich aus einer linearen Regression mit ihren eigenen Parametern, dem Gewicht a und dem Achsenabschnitt b, zusammen. Der optimale Zeitraum der für die Anwendung der linearen Regression zu

e | Schematische Darstellung der Daten verschiedener Bezirkskombinationen



wählenden Daten entsteht mit zusätzlichen Parametern (*Fenster*, *Fenster-Versatz* und *Meldewochen*), welche systematisch auf Basis des maximalen R^2 in historischen Daten gesucht werden. Bei festgelegtem Zeitraum werden, ebenfalls auf Basis des R^2, die besten Kombinationen der Bezirke gesucht. Die größte Herausforderung bei der Entwicklung dieses Verfahrens ist also die systematische Suche nach den optimalen Parametereinstellungen. Sie sind entscheidend, um die bestmögliche Regressionsschätzung berechnen zu können, die aus den vorhandenen Daten mit der geringen Anzahl an Merkmalen (Meldeverzug, Kalenderwoche und Bezirk) erreichbar ist.

Gleitende Mittelwerte als Alternative

Um herauszufinden, ob das bis zu diesem Punkt dargestellte Vorgehen einen Mehrwert gegenüber simpleren Herangehensweisen bringt, wurde zu Vergleichszwecken ein einfacher gleitender Mittelwert über die Zeitreihe gebildet. Dabei wird ein Mittelwert über die bereits bearbeiteten Sterbefallzahlen in einem vorab definierten Zeitraum berechnet und als Schätzwert für die aktuelle Woche verwendet. Dieses Vorgehen benötigt dementsprechend weniger Berechnungsschritte und Parameter als das Regressionsverfahren. Das Intervall, welches die Datenpunkte für die Mittelwertbildung liefert, definiert sich jedoch ebenfalls über die Parameter: *Fenster* (Länge in Wochen), *Fenster-Versatz* (in Wochen) und Anzahl *Meldewochen*. Der so entstandene Schätzwert schwankt naturgemäß relativ wenig. Die Bezirkskomponente ist kein Bestandteil der Berechnung des gleitenden Mittelwertes. Im Gegensatz zu Methoden des maschinellen Lernens werden derartige Glättungsverfahren nicht trainiert. Sie zählen zu den Werkzeugen der klassischen Zeitreihenanalyse. Für die hier untersuchte Fragestellung sind sie interessant, da sie in der Praxis ebenfalls für herkömmliche Prognosen verwendet werden.

Der größte Vorteil dieses Ansatzes liegt in der gering ausgeprägten Schwankung der Schätzwerte. Die Schätzwerte befinden sich relativ dicht an der zu schätzenden Zeitreihe und fluktuieren wenig. Im Gegensatz dazu sind die Ergebnisse des Regressionsansatzes stärkeren Schwankungen ausgesetzt und bilden gelegentlich Ausreißer. Andererseits zeigt die Mittelwert-Schätzung einen zeitlichen Versatz zur aktuellen Zeitreihe. Diese Trägheit ist der Art der Berechnung geschuldet und bedeutet, dass die Schätzungen nicht auf die aktuelle Situation eingehen können. Dies ist bei der Regressionsschätzung anders, da die Berechnungsvorschrift jeweils auch die aktuell bearbeiteten Fälle der zu schätzenden Woche berücksichtigt.

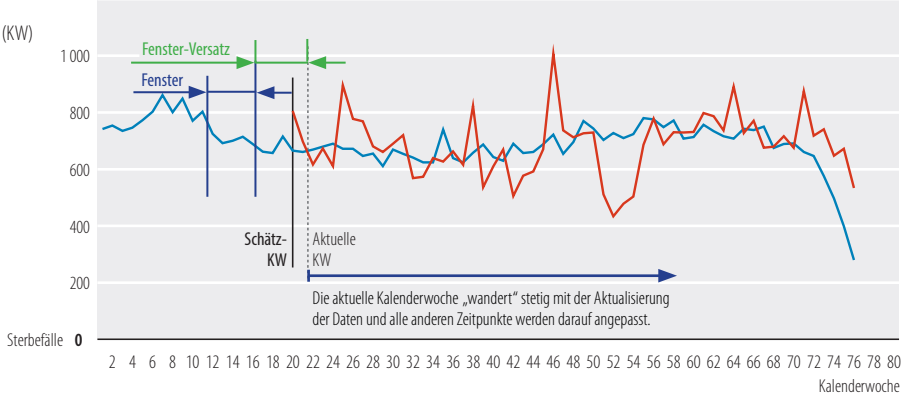
Da beide Verfahren tendenziell gegensätzliche Vor- und Nachteile besitzen, wurde eine dritte Methodik in die Analyse einbezogen: die Kombination beider Verfahren als einfacher Durchschnitt der beiden Schätzwerte. Diese einfache Idee ist der Hoffnung entsprungen, dass die Vorteile beider Verfahren verknüpft werden und dies eine noch bessere Schätzung nach sich zieht. Dieser Ansatz führt insbesondere bei wenigen *Meldewochen* zu besseren Schätzungen, allerdings hat es sich nicht pauschal als „die beste" Herangehensweise aller drei Methoden erwiesen.

Für unterschiedliche *Meldewochen* liefert jeweils ein anderes der drei hier vorgestellten Verfahren den besten Schätzwert. Für jede zu schätzende Woche der kürzeren Vergangenheit gewinnt entweder die lineare Regression, der gleitende Mittelwert oder die Kombination aus beiden Verfahren. In der praktischen Anwendung der Sterbefallschätzungen am aktuellen Rand werden dementsprechend alle Ansätze genutzt.

Modelle evaluieren

Um die beste Parameterkonstellation zu finden, muss das Ergebnis, welches unter Verwendung bestimmter Parameter entstanden ist, einer Bewertung unterzogen werden. Dafür werden die historischen Daten der letzten drei Jahre verwendet. Als Bewertungskriterium wurde die mittlere quadratische Abweichung (mean squared error, MSE) herangezogen. Diese Abweichung bestimmt, wie stark der geschätzte Wert um den zu schätzenden Wert streut. Sie ermöglicht die Angabe von Unsicherheitsintervallen – Intervalle, in denen sich der wahre Wert mit hoher Wahrscheinlichkeit befinden wird.

f | Beispiel einer Zeitreihe für die Anzahl an Sterbefällen in Berlin (tatsächliche Fallauszählung und Schätzung) beginnend mit Kalenderwoche 1 aus 2019 bis Kalenderwoche 26 aus 2020, fortlaufend nummeriert



Wie bereits erwähnt, wird während der Analyse auf gleitende Art und Weise für jede Kalenderwoche ein separater Schätzwert erstellt. Daraus entsteht eine Zeitreihe von Schätzwerten. Abbildung f zeigt eine solche Zeitreihe für die Anzahl der Sterbefälle in Berlin (tatsächliche Fallauszählung und Schätzung) beginnend mit Kalenderwoche 1 aus 2019 bis Kalenderwoche 26 aus 2020. Die Kalenderwochen sind dabei fortlaufend nummeriert. Für die geschätzte Zeitreihe wurde die Regressionsschätzung verwendet und die Länge des *Fensters* auf vier Wochen, der *Fenster-Versatz* auf fünf Wochen und die Anzahl der *Meldewochen* auf eine Woche gesetzt. Es wurden für alle Schätzungen der Zeitreihe ausschließlich Daten verwendet, die nach einer Woche registriert wurden, unabhängig davon, ob bereits mehr vorhanden waren. Daten, die in der zu schätzenden Kalenderwoche noch nicht hätten vorliegen können, werden ausgefiltert.

Das resultierende Ergebnis ist eine Zeitreihe, die ausschließlich Schätzwerte enthält (Abbildung f, rote Linie). Diese wird der wahren Zeitreihe (Abbildung f, blaue Linie) mittels MSE gegenübergestellt. Dabei liegt der Fokus darauf, wie stark sich die Schätzungen von den tatsächlich eingetretenen Sterbefällen unterscheiden – also: Was hat das Modell vorhergesagt und was ist tatsächlich eingetreten. Bei der Parametersuche werden sehr viele derartige Schätz-Zeitreihen generiert. Für jede beliebige Anzahl an *Meldewochen* kann auf diese Weise entschieden werden, wie *Fenster* und *Fenster-Versatz* gewählt werden müssen, sodass die geringste Abweichung zur wahren Zeitreihe eintritt. So kann die optimale Parameterkonstellation ermittelt werden. Damit die Parameterberechnungen nicht verfälscht werden, sind in der Analyse die Sterbefallmeldungen am aktuellen Rand bei der Bewertung mittels MSE ausgeschlossen. Diese Meldungen sind zum Zeitpunkt der Untersuchung tatsächlich nicht vollständig in der Datengrundlage enthalten. Damit fehlen die wahren Werte zum Vergleich.

Das Finden der Parameter verursacht einen relativ hohen Rechenaufwand. Sind die Parameter allerdings einmal gefunden, kann das Verfahren einfach angewendet werden. Der praktische Anwendungsteil verläuft prinzipiell ähnlich zur Analyse, nun jedoch für die unvollständigen zehn Wochen. Für diesen aktuellen Rand sind keine korrekten Output-Werte vorhanden. Es ist demzufolge (noch) keine Überprüfung möglich, ob die Schätzwerte die wahren Werte treffen.

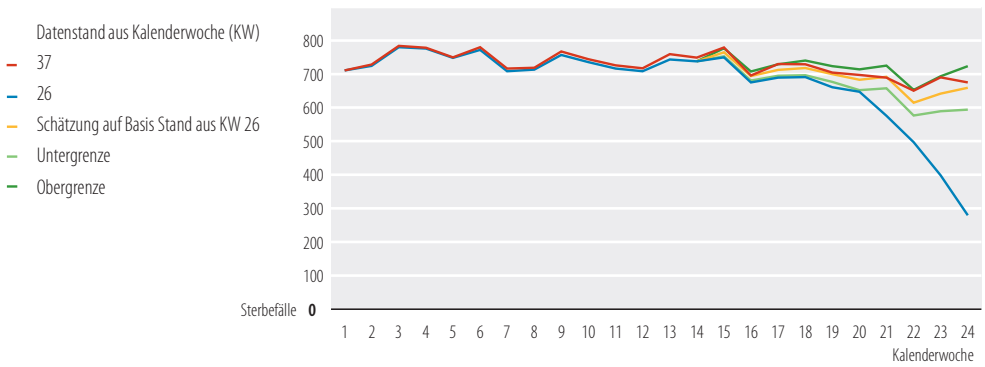
Hinsichtlich der Zuverlässigkeit des Verfahrens sollten zwei Aspekte berücksichtigt werden. Zum einen besteht generell eine Tendenz zur Unterschätzung. Die Sterbefallzahlen für Berlin werden kontinuierlich etwas zu gering eingeschätzt. Dieser Effekt entsteht dadurch, dass die für die Regression als Wahrheit angenommenen Daten häufig für einen sehr langen Zeithorizont Lücken enthalten. Vereinzelt werden Sterbefälle erst extrem spät gemeldet (nach 40 Wochen und später). Eine zu hohe Anzahl an Sterbefällen hingegen, die langfristig wieder nach unten korrigiert werden muss, tritt nicht auf. Zum anderen ist das vorgestellte Vorgehen nicht in der Lage, spontane Schwankungen des Meldeverhaltens abzubilden. Wie bereits erwähnt, werden möglichst konsistente Daten für das Zeitfenster, in dem die Regressionsgerade modelliert und angewendet wird, benötigt. In einem bestimmten Zeitraum sollte das Meldeverhalten entsprechend robust sein. Dies macht die Ergebnisse qualitativ hochwertig und verlässlich. Leichte Änderungen, die im Laufe der Zeit kontinuierlich entstehen, werden jedoch abgebildet.

Neben diesen Grenzen ist bei der Anwendung zu beachten, dass sogenannte „Corner-Cases“ in den zukünftigen Daten auftreten können. Dabei handelt es sich um unplausible Fälle oder eigenartige Konstellationen, die gesondert berücksichtigt werden müssen. So wurde beispielsweise in der 44. Kalenderwoche des Jahres 2017 kein einziger Sterbefall innerhalb der ersten Woche registriert. Dieser Umstand produziert bei der Modellierung der Regressionsgeraden für eine Meldewoche fehlende Werte, sodass ohne gesonderte Behandlung keine Regression erstellt werden kann. Wie mit solchen oder ähnlichen Ausnahmesituationen umgegangen werden soll, muss dann jeweils im Einzelfall genauer betrachtet werden. Neben diesen Sonderfällen sollten auch die bestimmten Parameter in regelmäßigen Abständen neu berechnet werden.

Ergebnis und Ausblick

Die praktische Anwendung des Verfahrens ist in Abbildung g zu sehen. Sie zeigt die Schätzung des aktuellen Rands auf Basis eines Datenstandes vom 23. Juni 2020 (Kalenderwoche 26). Die zu diesem

g | Sterbefälle (Schätzungen und Fallauszählungen) 2020 in Berlin nach Kalenderwochen



Zeitpunkt bekannten Sterbefälle sind blau eingezeichnet. In Gelb sind die darauf basierenden Schätzwerte für die bis dato zehn letzten Wochen zu sehen. Unvollendete Kalenderwochen werden nicht berücksichtigt: Der 23. Juni 2020 war ein Dienstag, sodass Kalenderwoche 26 ignoriert wird. Zusätzlich werden Schätzungen für mindestens eine *Meldewoche* zurück generiert. Demzufolge ist die erste Woche, für die eine Schätzung angegeben ist, Kalenderwoche 24. Im Vergleich zu den in der Einführung gezeigten Datenständen ist hier bereits eine kurzfristigere Betrachtung der aktuellen Vergangenheit möglich. Bei der Veröffentlichung der Fallauszählungen wurden vier Wochen per se nicht dargestellt (siehe Abbildung a).

Die gelb eingezeichneten Schätzwerte einzelner Kalenderwochen (Abbildung g) können mithilfe der drei verschiedenen Verfahren entstehen: Regression, gleitender Mittelwert und Durchschnitt beider Techniken. Ist beispielsweise aus der Analyse bekannt, dass der reine Regressionsansatz für acht *Meldewochen* zurück die beste Schätzung liefert, so wird für Kalenderwoche 17 die Regression verwendet. Bei vier *Meldewochen*, also der Schätzung für Kalenderwoche 21, erwies sich beispielsweise der Durchschnitt aus Regression und gleitendem Mittelwert als beste Methodik. Für eine *Meldewoche* (Kalenderwoche 24) könnte wiederum nur der gleitende Mittelwert das beste Schätzergebnis liefern. So kann es passieren, dass die jeweiligen Schätzwerte unterschiedlich berechnet werden. Letztendlich wird das Beste aus allen drei betrachteten Methoden angewendet. Ebenfalls angewendet werden die in der Analyse ermittelten MSE-Werte, um die grün abgebildeten Unsicherheitsintervalle zu berechnen. Im Rahmen einer regelmäßigen Durchführung der Analyse kann es dabei immer wieder zu leichten Änderungen kommen.

In der Gegenüberstellung zu der rot eingezeichneten Zeitreihe der tatsächlichen Sterbefälle (Datenstand vom 11. September 2020, Kalenderwoche 37) werden die Abweichungen erkennbar. Das Verfahren ist in der Lage, den generellen Verlauf der Sterbefälle gut zu schätzen. Insbesondere vor dem Hintergrund, dass die in blau eingezeichneten, stark abfallenden eingegangenen Fallauszählungen keine Interpretation erlauben, bietet die Schätzung eine gute erste Indikation für die Sterbefallzahlen am aktuellen Rand.

Zu sehen ist auch die genannte systematische Unterschätzung. Da bei der MSE-Berechnung diese Unterschätzung bereits bekannt ist, könnte in weiteren Untersuchungen geschaut werden, ob dieser Bias aktiv genutzt werden kann, um die eigentliche Schätzung weiter zu verbessern. Darüber hinaus können in zukünftigen Betrachtungen andere Verfahren des maschinellen Lernens gegenübergestellt werden. Eine denkbare Variante wäre beispielsweise die Anwendung von neuronalen Netzen. Die Herausforderung hier besteht in der verhältnismäßig geringen Anzahl verfügbarer Trainingsdaten. Methoden dieser Art benötigen in der Regel eine hohe Anzahl an zu trainierenden Gewichten. Es wäre spannend zu sehen, ob dennoch vergleichbare Ergebnisse erzielt werden können.

Kerstin Erfurth ist Referentin in der Stabsstelle *Querschnittsanalysen und Digitale Transformation* des Amtes für Statistik Berlin-Brandenburg.
Dr. Jörg Höhne leitet die Stabsstelle *Statistische Methoden und Grundsatzfragen* des Amtes für Statistik Berlin-Brandenburg.