Prognose von Wärmeverbräuchen

Matthias Dziubany[1], Jens Schneider[2], Anke Schmeink[3], Guido Dartmann[4], Klaus-Uwe Gollmer[5] und Stefan Naumann[6]

Abstract: Nicht nur die Datenaufbereitung und die Wahl eines passenden mathematischen Modells, sondern auch die Ermittlung der für die Prognose benötigten Features, sind für die Güte der Prognosen von hoher Relevanz. Am Beispiel einer Wärmeverbrauchsprognose für ein Gebäude des Umwelt- Campus Birkenfeld werden auftretende Probleme diskutiert und Lösungen vorgestellt. Im ersten Schritt wird jeweils mittels Linearer Regression und Neuronalen Netzen die Prognose für einen Folgetag erstellt. Anschließend wird die Güte der Prognose von Linearer Regression mit der von Neuronalen Netzen verglichen.

Keywords: Prognose; Regression; Neuronales Netz; Datenaufbereitung

1 Einleitung

In den meisten Fällen ist es nicht möglich, mathematische Modelle direkt ohne Vorverarbeitung auf Rohdaten anzuwenden. Zudem kann eine ungünstige Wahl der betrachteten Merkmale (Features) schlechte Ergebnisse liefern. Im Folgenden wird der Prozess von der Rohdatenvorverarbeitung bis hin zur Anwendung eines Prognosemodells am Beispiel einer Wärmeverbrauchsprognose für ein Nichtwohngebäude des Hochschulstandorts Umwelt- Campus Birkenfeld aufgezeigt.

Solche Analyse- und Prognoseverfahren helfen zum einen, die Qualität von Datenbeständen besser zu beurteilen, und können zum anderen perspektivisch das Smart Grid, also die enge Verschränkung von Erzeugern und Verbrauchern im Strombereich auch im Wärmebereich adressieren. Diese genauere Kopplung, welche ebenfalls kleinteilige Vorhersagen umfasst und so eine bedarfsgerechtere Produktion und bei entsprechender Planung sogar eine unmittelbare Abnahme von erzeugter Wärme oder erzeugtem Strom ermöglicht, ist ein wichtiger Beitrag zur sogenannten Energiewende, da es beispielsweise die Speicherbedarfe reduziert und damit Verluste minimiert. Idealerweise kommen so Prognosen für die Verbrauchsseite und Prognosen für die Strom-/Wärmerzeugung zusammen.

[1] Hochschule Trier, Campusallee, 55768 Hoppstädten-Weiersbach, Germany, m.dziubany@umwelt-campus.de
[2] Hochschule Trier, Campusallee, 55768 Hoppstädten-Weiersbach, Germany, j.schneider@umwelt-campus.de
[3] RWTH Aachen, Kopernikusstr. 16, 52074 Aachen, Germany, anke.schmeink@rwth-aachen.de
[4] Hochschule Trier, Campusallee, 55768 Hoppstädten-Weiersbach, Germany, g.dartmann@umwelt-campus.de
[5] Hochschule Trier, Campusallee, 55768 Hoppstädten-Weiersbach, Germany, k.gollmer@umwelt-campus.de
[6] Hochschule Trier, Campusallee, 55768 Hoppstädten-Weiersbach, Germany, s.naumann@umwelt-campus.de

2 Datenherkunft

Für die Erstellung der im Folgenden beschriebenen Prognose werden neben dem Wärmeverbrauch die Innentemperaturen einzelner Räume, die Außentemperatur und der Wert der Globalstrahlung betrachtet. Dass sämtliche dieser Werte für einen längeren Zeitraum vorliegen, ist keine Selbstverständlichkeit. In einem vorhergegangenen Forschungsprojekt REGENA [Na16] wurden über dessen Verlauf für Teile des Umwelt-Campus Birkenfeld eine Vielzahl an Daten, darunter auch der Heizenergieverbrauch für einzelne Gebäude und die Raumtemperaturen einzelner Räume, erfasst. Die Außentemperatur und die Globalstrahlung wurden von der lokalen Wetterstation bezogen, wodurch eine genaue Geolokalisierung der Werte gegeben ist.

3 Datenaufbereitung

Bereits bei der Zusammenstellung der benötigten Daten können an einigen Stellen Schwierigkeiten auftreten. Insbesondere wenn die Messdaten nicht explizit mit dem Ziel der Erstellung einer Verbrauchsprognose erfasst und organisiert wurden, müssen zunächst die benötigten Datensätze identifiziert werden. Dabei müssen unter Umständen auch bautechnische Gegebenheiten berücksichtigt werden, beispielsweise die Klärung der Frage, welche Messstellen zu welchen Räumen bzw. Gebäuden gehören.

Eine weitere Herausforderung ist der Umgang mit fehlenden Daten. Die einfachste Vorgehensweise ist die Entfernung jedes Samples mit mindestens einem fehlenden Feature aus dem Datensatz, was jedoch, je nach Anzahl und Verteilung der fehlenden Daten, den Verlust einer großen Menge an Samples bedeuten kann. Alternativ kann versucht werden, fehlende Daten durch Interpolation oder maschinelle Lernmethoden zu rekonstruieren.

In der beschriebenen Prognose wurde eine Rekonstruktion jedoch nicht in Betracht gezogen, da sich die fehlenden Daten teilweise über große, zusammenhängende Perioden erstreckten. Stattdessen wurde auch bei der Auswahl der Features auf deren Vollständigkeit bezüglich der Messwerte geachtet.

Die Auswahl und Anzahl von betrachteten Features ist maßgeblich für die Komplexität der Erstellung eines Prognosemodells und dessen Qualität. Da einige Modelle Probleme mit sehr hoch dimensionalen Eingabedaten haben oder dann eine sehr große Anzahl an Samples benötigen, sollte auch die Anzahl der verwendeten Features, wenn möglich, gering gehalten werden. Ein einfacher Ansatz zur Reduktion der Dimensionalität von Features ist zum Beispiel die Bildung von Mittelwerten. Zur Erstellung der Prognose wurden verschiedene Kombinationen der Features getestet. Infolgedessen wurden einige Features, die ursprünglich zur Verwendung vorgesehen waren, wegen mangelnder Relevanz (Globalstrahlung) oder zu vieler fehlender Daten (Raumtemperaturen) verworfen.

Die letztendlich verwendeten Features sind die Wärmeverbräuche $Q(t)$ und die gemittelten Außentemperaturen $T(t)$ der letzten 14 Tage. Zusätzlich wurden eine Saisonkomponente $S(t)$ und ein binärer Werktagvektor $W(t)$ verwendet. Die Saisonkomponente ist ein häufig verwendeter Indikator [TT17] für die Jahreszeit und hat am 1.1. und am 31.12. den Wert 1 und am 2.7. den Wert -1. Sie berechnet sich wie folgt:

$$S(t) = \cos\left(\frac{2\pi}{365} \cdot t\right), t \in \{1, \dots, 365\}.$$

Der Werktagindikator gibt an, ob es sich bei dem zu prognostizierenden Tag um einen Werk- (0) oder Feiertag (1) handelt.

Die folgende Tabelle zeigt eine Übersicht über die für die Prognose vorbereiteten Features.

	Vortagesverbrauch			**Vortagestemperatur**			**Saison komponente**	**Werk-tag**
	1	**…**	**14**	**1**	**…**	**14**		
18.04.2013	0,2170	…	0,5890	16,0594	…	3,1189	-0,2844	0
19.04.2013	0,2690	…	0,6760	12,1705	…	0,3735	-0,3008	0
20.04.2013	0,3510	…	0,5290	8,4966	…	1,5584	-0,3172	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
29.04.2016	0,4450	…	0,2979	3,5902	…	10,1240	-0,4750	0

Tab. 1 Einzelne Features in den Spalten und den Samples mit Zeitstempeln in den Zeilen

4 Prognose für den Folgetag

Um die Güte der Prognosemodelle zu bestimmen, werden die Samples in Trainings- und Testdaten aufgeteilt. Die Trainingsmenge ergibt sich aus m=400 Samples, welche vom 18.04.2013 bis zum 28.04.2015 reichen und die Testmenge aus n=280 Samples, welche vom 29.04.2015 bis zum 29.04.2016 reichen. Aufgrund von Messfehler und fehlenden Daten ist nicht für jeden Tag ein Sample vorhanden. Die Wärmeverbräuche der übrig gebliebenen Tage und die Aufteilung in Trainings- und Testmenge ist in Abbildung 1 zu sehen.

Mit dieser Trainingsmenge wurden durch Lineare Regression bzw. Neuronale Netze Prognosen für den Wärmeverbrauch des Folgetags erstellt. Hervorzuheben ist, dass die Daten aus dem Testjahr vom 29.04.2015 bis zum 29.04.2016 nicht im Training berücksichtigt werden und die Testmenge somit besonders gut für die Bestimmung der Güte von zukünftigen Prognosen geeignet ist.

Im Folgenden wird die Güte der Prognosen stets auf der Testmenge anhand der im nächsten Kapitel beschriebenen Bewertungsmaße bestimmt und wie in [TT17] durch Streudiagramme, Histogramme und Zeitverläufe visualisiert.

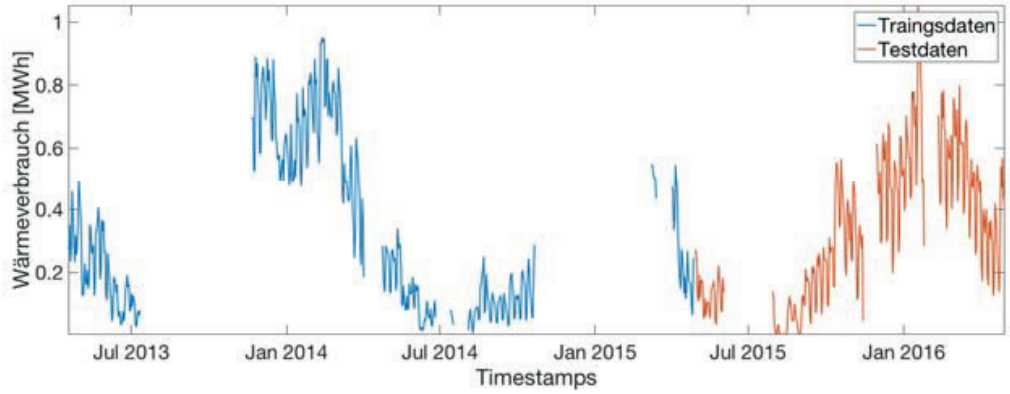

Abb. 1: Zeitlicher Verlauf der Wärmeverbräuche und die Aufteilung in Trainingsdaten (blau) und Testdaten (rot)

4.1 Bewertungsmaße

Zur Bestimmung der Güte der verschiedenen Prognosen werden der RMSE (Root Mean Square Error), der Determinationskoeffizient für multiple lineare Regression r^2 und der SMAPE (Symmetric Mean Absolute Percentage Error) verwendet [HK06]. Der RMSE berechnet die Wurzel des MSE (Mean Square Error), welcher das arithmetische Mittel der Fehlerquadrate bestimmt. Es sei n gleich der Anzahl der Samples in der Testmenge, $y(t)$ der wahre Messwert und $\hat{y}(t)$ der vorhergesagte Wert, dann ist der RMSE durch folgende Formel gegeben:

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(y(t) - \hat{y}(t))}{n}}$$

Der Determinationskoeffizient ist ein normiertes Maß und berechnet die Gesamtvariation als das Verhältnis der durch die Regression gegebenen Variation und der zu erklärenden Variation. Für die Heizenergieprognose wird das Bestimmtheitsmaß für multiple Lineare Regression verwendet. Dieses ist durch folgende Formel gegeben:

$$r^2 = \frac{[\sum_{t=1}^{n}(y(t)-\bar{y}(t))\cdot(\hat{y}(t)-\bar{\hat{y}}(t))]^2}{[\sum_{t=1}^{n}(y(t)-\bar{y}(t))^2]\cdot[\sum_{t=1}^{n}(\hat{y}(t)-\bar{\hat{y}}(t))^2]},$$

wobei $\bar{y} = \frac{1}{n}\sum_{t=1}^{n} y(t)$ und $\bar{\hat{y}} = \frac{1}{n}\sum_{t=1}^{n} \hat{y}(t)$ ist.

Der SMAPE ist ein normiertes Gütemaß welches auf prozentualen Fehlern basiert und sich wie folgt berechnet:

$$SMAPE = \frac{100\%}{n} \sum_{t=1}^{n} \frac{|y(t) - \hat{y}(t)|}{((|y(t)| + |\hat{y}(t)|) \div 2)}$$

Die Betrachtung aller drei Gütemaße ermöglicht eine gute Einschätzung der Güte der Prognose.

4.2 Lineare Regression

Wenn alle Features einen linearen Einfluss auf den Wärmeverbrauch besitzen, eignet sich eine Lineare Regression hervorragend als Prognosemodell. Die linearen Einflüsse der Features werden durch folgende multiple lineare Modellfunktion beschrieben:

$$\begin{aligned} Q(t+1) = x(t) \cdot p = p\_0 \; & + p\_1 \cdot Q(t) + p\_2 \cdot Q(t-1) + \cdots + p\_14 \cdot Q(t-13) \\ & + p_{15} \cdot T(t) + p_{16} \cdot T(t-1) + \cdots + p_{28} \cdot T(t-13) \\ & + p_{29} \cdot S(t+1) \\ & + p_{30} \cdot W(t+1) \end{aligned}$$

wobei folgende Notationen gelten:

p $= [p_0, \dots, p_{30}]$,
$x(t)$ $= [1, Q(t), \dots, Q(t-13), T(t), \dots, T(t-13), S(t+1), W(t+1)]$,
$Q(t)$ $=$ Wärmeverbrauch von Tag t,
$T(t)$ $=$ Außentemperatur von Tag t,
$S(t)$ $=$ Saisonkomponente von Tag t,
$W(t)$ $=$ Werktagindikator von Tag t.

Da Prognosemodelle im Allgemeinen nicht lineare Einflüsse besitzen und Daten Messfehler behaftet sind, ist es nicht möglich, p so zu wählen, dass in der Modellfunktion für alle t Gleichheit gilt. Aus diesem Grund wird p so gewählt, dass $\sum_{t=1}^{m} (Q(t+1) - x(t) \cdot p)^2$ minimal ist.

Es sei $X = \begin{bmatrix} x(1) \\ \vdots \\ x(m) \end{bmatrix} \in \mathbb{R}^{m \times 30}$ und $Q = \begin{bmatrix} Q(1+1) \\ \vdots \\ Q(m+1) \end{bmatrix} \in \mathbb{R}^{m}$.

Dann kann das optimale p^* durch die Moore-Penrose Pseudoinverse $X^{\#}$, die in Matlab mit dem Befehl $pinv(X)$ bestimmt werden kann, wie folgt berechnet werden:

$$p^* = X^{\#} \cdot Q$$

Um die Güte des aufgestellten Modells zu bestimmen, wird die Prognose eines Testdatums $x(t) \cdot p^*$ dem wahren Wert $Q(t+1)$ gegenüberstellt. Somit ergeben sich für die betrachteten Gütemaße folgende Werte:

RMSE	r^2	SMAPE
0,0687 [MHz]	0,9093	11,6235 [%]

Tab. 2 Güte der Linearen Regression

Abbildung 2 stellt die durch Lineare Regression prognostizierten Wärmeverbräuche den tatsächlichen Verbräuchen als Streudiagramm und als Histogramm gegenüber. Liegt ein Wert im Streudiagramm auf der Geraden mit Steigung 1, entspricht die Prognose genau dem wahren Wert.

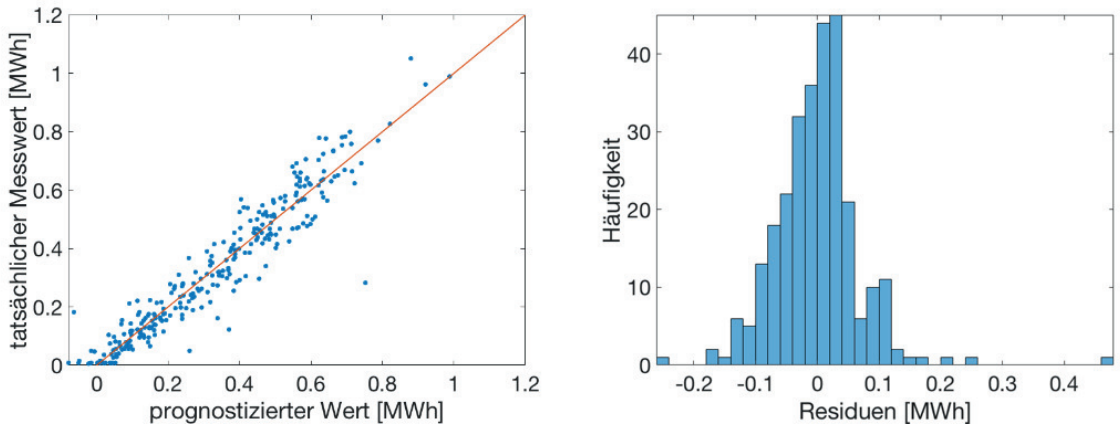

Abb. 2: Streudiagramm (links) und Histogramm (rechts) der durch Lineare Regression prognostizierten Wärmeverbräuche im Vergleich mit den tatsächlichen Messwerten aus der Testmenge

In Abbildung 3 ist der zeitliche Verlauf der Folgetagsprognosen durch Lineare Regression und der tatsächlichen Verbräuche über das Testjahr zu sehen. Lücken im Graphen resultieren aus fehlenden Messwerten.

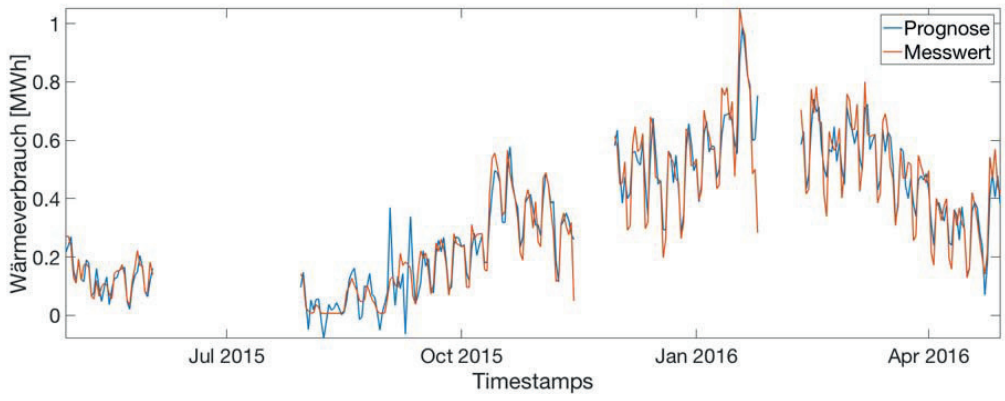

Abb. 3: Zeitlicher Verlauf der Folgetagsprognosen durch Lineare Regression und der tatsächlichen Verbräuche über das Testjahr

4.3 Neuronale Netze

Im Gegensatz zur Linearen Regression kann ein neuronales Netz auch nicht lineare Zusammenhänge abbilden, stellt dafür jedoch bezüglich des exakten Zustandekommens der Prognose eine Black Box dar.

Grundsätzlich besteht ein Neuronales Netz aus einem Input Layer, einer Anzahl von Hidden Layern und einem Output Layer. Die Anzahl der Units (Neuronen) im Input Layer entspricht der Anzahl der betrachteten Features und für jeden vorherzusagenden Output gibt es eine Unit im Output Layer. Die Konfiguration der Hidden Layer, das heißt die Anzahl der Layer und der Units pro Layer, kann frei gewählt werden, hat jedoch Einfluss auf die Qualität des Netzes.

Units verschiedener Layer können beliebig durch gerichtete gewichtete Kanten verbunden werden. Jede Unit besitzt eine Aktivierungsfunktion, die bezüglich den Gewichten der eingehenden Kanten und den Outputs der verbundenen Units einen eigenen Output erzeugt. Die Kantengewichte w_{ij} der einzelnen Kanten werden nach einer zufälligen Initialisierung in einem Trainingsprozess berechnet. Dabei werden Trainingsdaten mit bekannten Ergebnissen in das Netz gegeben und die Gewichte derart angepasst, dass der Fehler zwischen den bekannten Ergebnissen und den Outputs des Netzes minimal wird. Dabei ist zu beachten, dass bei der Minimierung des Fehlers möglicherweise nur ein lokales Minimum erreicht wird und die zufällige Initialisierung zu verschiedenen Endgewichten führen kann.

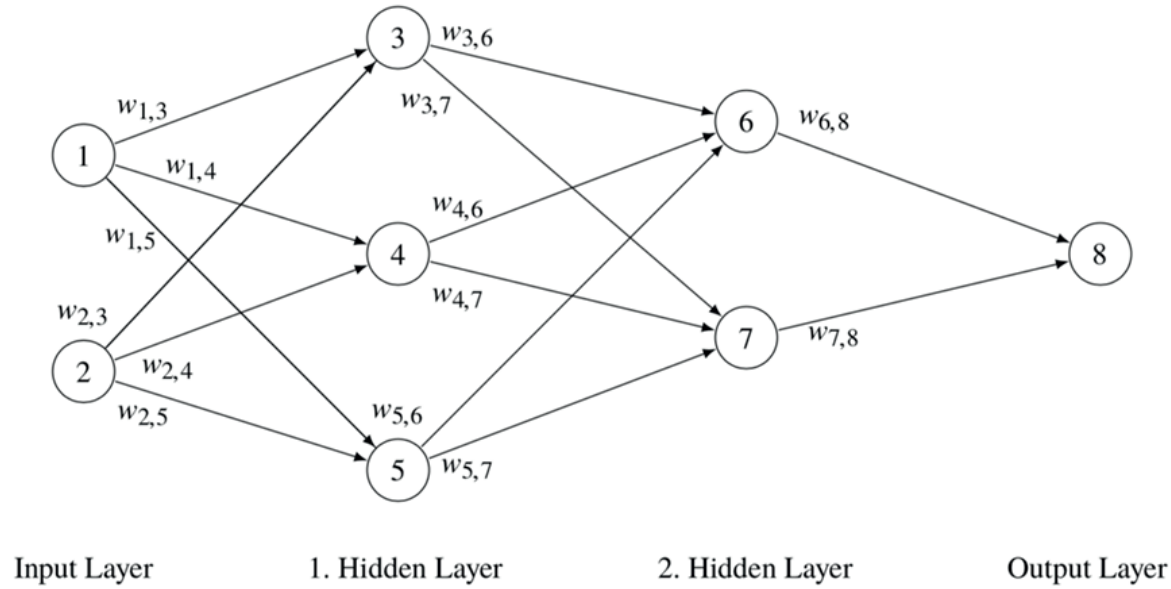

Abb. 4: Neuronales Netz mit 2 Input Units, 3 Neuronen im ersten Hidden Layer, 2 Neuronen im zweiten Hidden Layer und einer Output Unit

Für die hier beschriebene Prognose werden Feed Forward Netze mit zwei Hidden Layern verwendet. Feed Forward Netze sind vollständig vernetzt, das heißt jede Unit eines

Layers besitzt eine gerichtete Kante zu jeder Unit des folgenden Layers. Ein mögliches Feed Forward Netz mit zwei Input Units, einer Output Unit und drei Neuronen im ersten Hidden Layer und zwei im zweiten ist in Abbildung 4 zu sehen.

Für die Prognose des Wärmeverbrauchs mit der zuvor beschriebenen Feature Auswahl hat das Netz insgesamt 30 Input Units (je 14 für die Wärmeverbräuche und Außentemperaturen der letzten zwei Wochen, die Saisonkomponente und den Werktagindikator) und genau eine Output Unit die den prognostizierten Verbrauch ausgibt. Eine passende Anzahl an Units pro Hidden Layer wird heuristisch bestimmt, indem verschiedene Kombinationen mehrfach getestet werden und per Mittelwertbildung die Beste ausgewählt wird.

Das so bestimmte Netz mit 3 Units im ersten und 3 Units im zweiten Hidden Layer hat angewandt auf die definierte Testmenge folgende Güte:

RMSE	r^2	**SMAPE**
0,0634 [MHz]	0,9248	11,6116 [%]

Tab. 3 Güte des Neuronalen Netzes

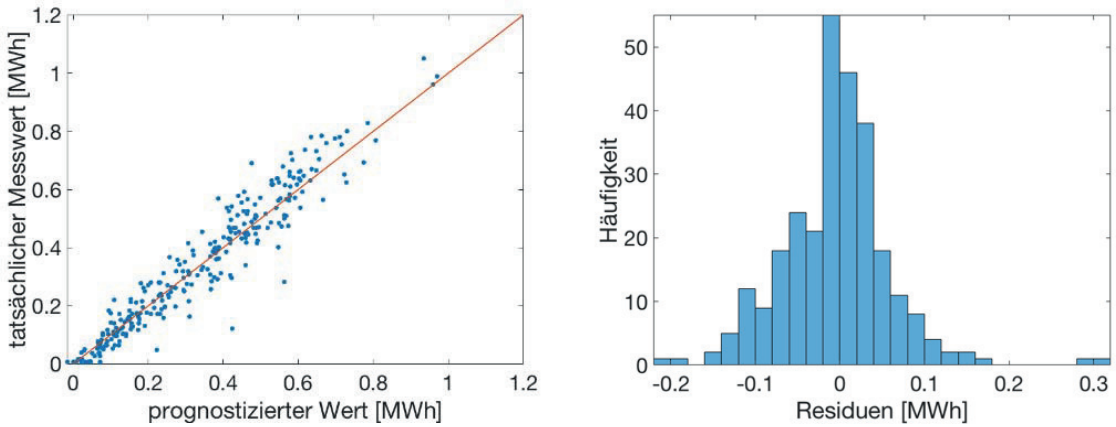

Abb. 5: Streudiagramm (links) und Histogramm (rechts) der durch das Neuronale Netz prognostizierten Wärmeverbräuche im Vergleich mit den tatsächlichen Messwerten aus der Testmenge

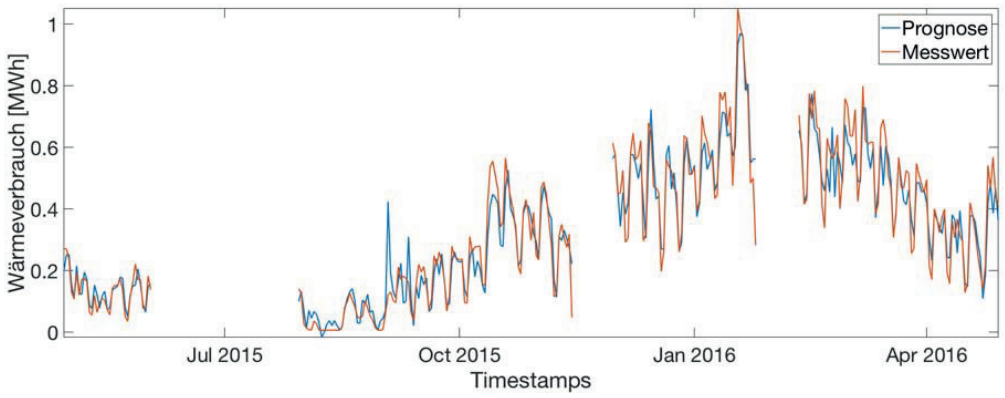

Abb. 6: Zeitlicher Verlauf der Folgetagsprognose durch das Neuronale Netz und der tatsächlichen Verbräuche über das Testjahr

5 Fazit

Die Wahl der Features beeinflussen die Komplexität und die Güte der Prognose erheblich. Bereits die Betrachtung der Vortagesverbräuche und der Vortagesaußentemperaturen, sowie einer Saisonkomponente und eines Werktagindikators führen auf unseren Daten zu einer ausreichend guten Prognose. Diese kann durch weitere Trainingsdaten verbessert werden. Der Vergleich beider Modelle in Tabelle 4 zeigt außerdem, dass Neuronale Netze und Lineare Regression mit den gewählten Features eine ähnliche Genauigkeit erreichen. Dies gilt allerdings unter dem Vorbehalt, dass das Neuronale Netz so gewählt wurde, dass es die "beste" Güte auf der Testmenge besitzt. Folglich ist die Lineare Regression aufgrund dieser Tatsache und ihrer geringeren Komplexität bei der hier betrachteten Datenmenge besser geeignet.

	Lineare Regression	**Neuronales Netz**
RMSE [MWh]	0,0687	0,0634
r^2	0,9093	0,9248
SMAPE [%]	11,6235	11,6116

Tab. 4 Vergleich Lineare Regression mit Neuronalen Netzen anhand von RMSE, r^2 und SMAPE

Da Wärmeverbrauchsprognosen auch für mehrere Folgetage benötigt werden, sollte diese Prognose erweitert werden. Hierzu könnten auch Wettervorhersagen die Güte der Prognose verbessern.

Danksagung

Dieses Projekt wurde vom Bundesministerium für Bildung und Forschung (BMBF) unter dem Förderkennzeichen 01IS17073 gefördert.

Literatur

[Bi09] Bishop, C.M.: Pattern recognition and machine learning. Springer, New York, NY, 2009.

[HK06] Hyndman, R.J.; Koehler, A.B.: Another look at measures of forecast accuracy. International Journal of Forecasting/, S.679-688, 2006.

[HTF08] Hastie, T.; Tibshirani, R; Friedman, J.: The elements of statistical learning, Data mining, inference, and prediction. Springer, New York, NY, 2008.

[Na16] Naumann, S.; Christian, A.; Göttert, C.; Gollmer, K.-U.; Michels, R.; Rüffler, S.: Energieeinsparung im Gebäudebetrieb durch visualisiertes Feedback an Nutzer: Datenerfassung und Datenvisualisierung in Nicht-Wohngebäuden. In (Mayr, H.C.; Pinzger, M., Hrsg.): Informatik 2016. Gesellschaft für Informatik e.V., Bonn, S. 1239-1249, 2016.

[TT17] Tritschler, M.; Trischtler, M.: Monitoring und Betriebsoptimierung – Vergleich der Prognose des Energieverbrauchs mit neuronalen Netzen und linearen Modellen. GI – Gebäudetechnik in Wissenschaft & Praxis 138/04, S. 294-303, 2017.