Datenanalyse und Prozessoptimierung für Kanalnetze und Kläranlagen mit CI-Methoden

T. Bartz-Beielstein, M. Bongards, C. Claes, W. Konen, H. Westenberger

Fakultät für Informatik und Ingenieurwissenschaften
Fachhochschule Köln
Tel.: (02261) 8196-0
Fax: (02261) 8196-15
E-Mail: thomas.bartz-beielstein@fh-koeln.de

Zusammenfassung

Prozesse optimal zu steuern und Prognosen über ihren zukünftigen Verlauf anzustellen, gehört zu den wichtigsten, aber auch schwierigsten Aufgaben der industriellen Praxis. Wir demonstrieren, wie aktuelle Methoden der Computational Intelligence (CI) und des Data Mining ebenso wie klassische Ansätze zur Modellierung, Simulation und Optimierung von Kanalnetzen und Kläranlagen eingesetzt werden können. Dabei zeigt sich, dass die Kombination von Standardverfahren aus der Zeitreihenanalyse mit dem Verfahren der sequentiellen Parameteroptimierung schnell zu problemspezifischen Vorhersagemodellen führen kann.

1 Modellierung von Füllständen in Regenüberlaufbecken

Böden und Kanalnetze stellen ein komplexes dynamisches System dar. Offensichtlich haben die aktuellen Zustände der Böden einen wichtigen Einfluss auf den Füllstand der Überlaufbecken, da z.B. trockene Böden ein anderes Abflussverhalten zeigen als feuchte oder gar gefrorene Böden. Hinzu kommen noch weitere Einflussfaktoren wie die Sonnenscheindauer, die landwirtschaftliche Nutzung, Temperatur usw. Im Forschungsprojekt KANNST (KANalNetz-Steuerung) wird die Modellierung und Prognose von Füllstandshöhen in Regenüberlaufbecken auf Basis einzelner Regenmessungen untersucht [1]. Die Modellierung beruht auf empirisch erhobenen Regendaten, die vom Aggerverband zur Verfügung gestellt wurden. Der Aggerverband ist ein Wasserverband in Nordrhein-Westfalen. In seinem Verbandsgebiet übernimmt er für seine Mitglieder (Kreise, Städte, Gemeinden, Industrie und öffentliche Wasserversorgungsunternehmen) alle Aufgaben der Wasserwirtschaft. Die Regendaten dienten als Eingabedaten für ein Kanalnetzsimulationsprogramm. Wir wählten hierzu das Storm Water Management Model (SWMM) [2]. SWMM wird bereits seit 1971 weltweit zur Simulation von Kanalnetzen eingesetzt.

Die Regendaten wurden minütlich über einen Zeitraum von 108 Tagen erhoben, so dass insgesamt $n = 155.521$ Datensätze zur Verfügung standen. Untersuchungen mit anderen Datensätzen wurden zusätzlich durchführt, da von Anwenderseite ein großes Interesse an diesen Prognosen besteht. Basierend auf den gemessenen Regendaten x_t ($t = 1, 2, \ldots, n$), die im Folgenden als „Rain data“ bezeichnet werden, sollen die simulierten Füllstandshöhen in den Becken möglichst genau vorhergesagt werden. Die simulierten Werte y_t werden als „Outflow (target)“ und die von unseren Modellen vorhergesagten Werte $\hat{y}_t$ werden als „Outflow (prediction)“ bezeichnet. Bekannt sind somit x_t und y_t, zu bestimmen ist ein

Modell zur Berechnung von $\hat{y}_t$, so dass der Fehler zwischen y_t und $\hat{y}_t$ möglichst gering ist.

Unterschiedliche Ansätze wie *neuronale Netze* (NN), *Echo State Networks* (ESN), Differentialgleichungen oder die Modellierung mit Integralgleichungen werden dabei von uns eingesetzt [3].

Im zweiten Abschnitt stellen wir das Vorhersagemodell vor und beschreiben die zugehörigen Schritte zur Datenvor- und nachbearbeitung. Ebenfalls in diesem Abschnitt wird dargelegt, wie die Parameter des von uns favorisierten Modells eingestellt wurden. Im letzten Abschnitt geben wir eine kurze Zusammenfassung.

2 Verfahren

2.1 Das Vorhersagemodell

Erste Ansätze mit NN brachten nicht die gewünschten Ergebnisse. Der Einsatz von FIR-Filtern (Filter mit endlicher Impulsantwort, engl.: „finite impulse response filter“) führte mit geringem Modellierungsaufwand zu guten Ergebnissen. FIR-Filter verfügen über eine Impulsantwort mit garantiert endlicher Länge. Daher können FIR-Filter, unabhängig davon, wie die Filterparameter gewählt werden, niemals instabil werden oder zu einer selbstständigen Schwingung angeregt werden.

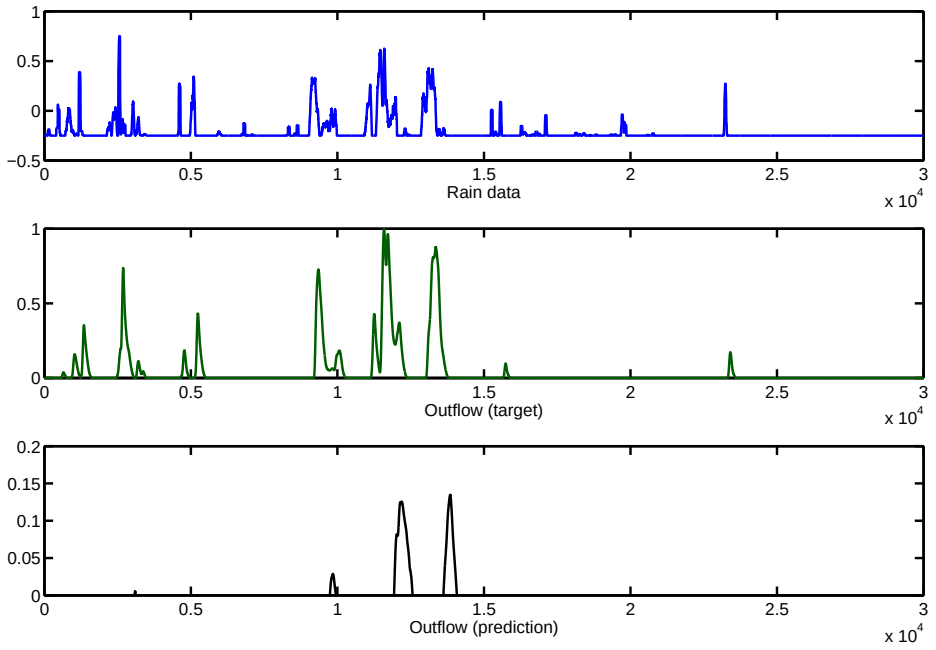

Bild 1: Vorhersage mit „intuitiv“ gewählten Modellparametern. Oben: Gemessener Niederschlag x_t. Mitte: Simulierter Abfluss y_t. Unten: Durch das FIR-Filter Modell vorhergesagter Abfluss $\hat{y}_t$.

In unserem Fall war die einfache Anwendung eines FIR-Filters nicht ausreichend. Eine passende Datenvor- und Nachbearbeitung war, ebenso wie eine Einstellung der Filterparameter, erforderlich. Die Regendaten wurden in vier Schritten verarbeitet: Normierung, Translation, Glättung und Skalierung.

Normierung: Im ersten Schritt wurden die Regendaten x_t auf das Intervall $[0; 1]$ abgebildet.

Translation: Im zweiten Schritt wurde die Verdunstung des Niederschlags modelliert. Von den gemessenen und normierten Werten wurde ein konstanter Betrag s abgezogen.

Exponentielle Glättung: Im dritten Schritt wurden l Gewichte c_i berechnet, die nach einer gewissen Verzögerung d (*delay*) mit wachsendem i exponentiell abnahmen. Hierzu war ein Abklingfaktor a zu bestimmen. Mit den so berechneten Gewichten wurden mit dem FIR-Filter die normierten und translatierten Regendaten geglättet.

Skalierung: Die geglätteten Werte wurden mit dem Faktor s_2 skaliert und negative Vorhersagewerte wurden auf Null gesetzt.

Die ersten Modellierungsversuche mit den so angepassten FIR-Filtern verliefen vielversprechend, so dass dieser Ansatz weiter verfolgt wurde. Abbildung 1 zeigt, dass einzelne Simulationswerte tendenziell erfasst werden (z.B. im Bereich von 70.000 bis 150.000), andere Bereiche jedoch überhaupt nicht.

Für das aus der exponentiellen Glättung und der Datenvorverarbeitung bestehende Vorhersagemodell sind fünf Parameter zu bestimmen. In vielen Situationen wird an dieser Stelle versucht, manuell, d.h. durch Probieren, günstige Parametereinstellungen für das Modell zu finden. Diese Vorgehenweise hat in vielerlei Hinsicht große Nachteile im Vergleich zu einer systematischen Vorgehensweise. Im Bereich der Simulation wurde dies von Kleijnen ausführlich dargestellt [4]. Wir setzten die sog. *sequentielle Parameteroptimierung* (SPO) ein [5].

2.2 Sequentielle Parameteroptimierung

SPO wurde speziell für Optimierungsprobleme entwickelt und kombiniert Ansätze aus der klassischen Regression und Varianzanalyse mit modernen statistischen Verfahren. [6] geben eine kurze Einführung in die sequentielle Parameteroptimierung. SPO verwendet eine sequentielle Vorgehensweise zur Bestimmung guter Parametereinstellungen. Zudem werden statistische Daten generiert, die während oder nach Abschluss der Optimierung wertvolle Informationen über das FIR-Filter-Modell liefern. Interessant ist z.B. die Analyse, welcher der fünf Faktoren den größten Einfluss auf die Vorhersagegüte hat.

2.2.1 Region of Interest

SPO erfordert die Spezifikation des Suchraums, der sog. *region of interest*, für mögliche Parametereinstellungen. Erfahrungswerte für gute Einstellungen lagen in unserem Fall nicht vor, da dieses Modell zum ersten Mal für diese Daten herangezogen wurde. Es gab aber Bereiche, die für das Modell als „sinnvoll“ zu bezeichnen waren. So waren zu große Werte für die Verzögerung (d) nicht sinnvoll, wohingegen die Auswirkung kleiner Werte noch zu untersuchen war. Es war zu Beginn der Untersuchungen nicht klar, ob die Verzögerung überhaupt einen positiven Effekt auf die Modellierungsgüte hat. Für die Verzögerung wählten wir daher den Bereich von 0 bis 1000 als *region of interest*. Die Translation s wurde im Bereich von 0 bis 1 variiert. Für alle anderen Parameter wurde der Bereich von 1 bis 1000 gewählt. Diese Einstellungen sind in Tabelle 1 zusammengestellt.

Tabelle 1: Region of interest für die Parameter des Vorhersagemodells.

Parameter	Abkürzung	Symbol	untere Grenze	obere Grenze
Verdunstung	SHIFT	s	0	1
Anz. Gewichte	LENGTH	l	1	1000
Abklingfaktor	EXP	a	1	1000
Verzögerung	DELAY	d	1	1000
Skalierung	SCALE2	s_2	1	1000

2.2.2 Gütefunktion

Zur Güte der berechneten Vorhersage wurde die Summe der quadratischen Abstände $\sum_1^n (y_t - \hat{y}_t)^2$ zwischen simulierten und vorhergesagten Werten berechnet, so dass ein Minimierungsproblem vorlag.

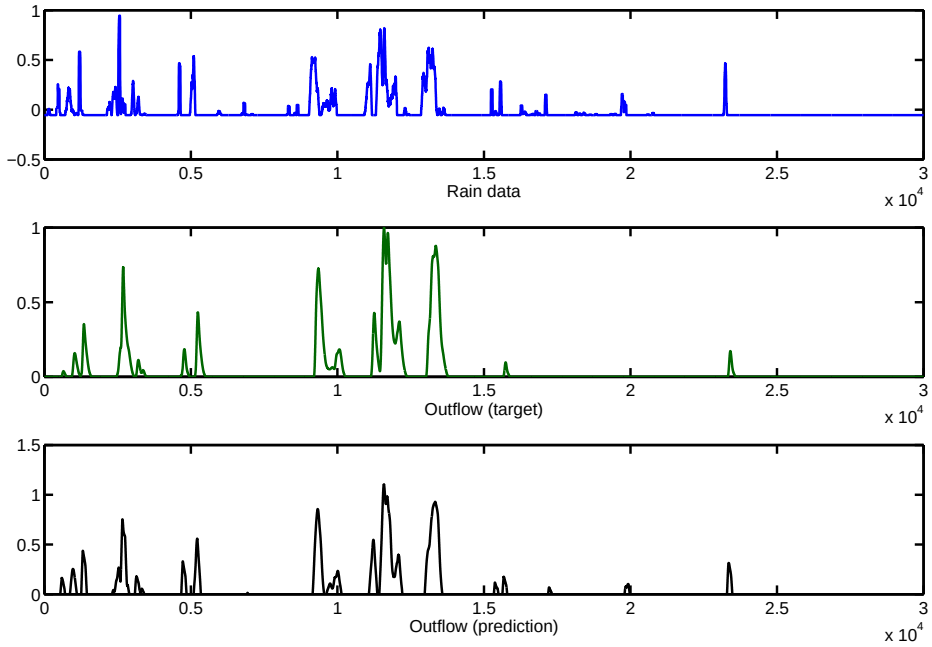

Bild 2: Vorhersage mit dem durch SPO verbesserten Modell. Es wurden 50 SPO Schritte durchgeführt. Die gesamte Optimierung benötigt auf einem Standard-PC (2GHz, 2GB RAM) nur wenige Minuten.

2.2.3 Ergebnis der Optimierung mit SPO

Mit der sequentiellen Parameteroptimierung konnte der Fehler auf 0,000248612 reduziert werden. Dieses Ergebnis wurde mit den Einstellungen $s = 0,055576$, $l = 154$, $a = 212,12$, $d = 78$ und $b = 61,62$ erzielt. Die auf diesen Einstellungen basierende Vorhersage ist in Abbildung 2 dargestellt.

Während eines SPO Laufs kann der Einfluss und die Veränderung der einzelnen Parameter im FIR-Filter-Modell visualisiert werden. In Abbildung 3 werden diese Informationen

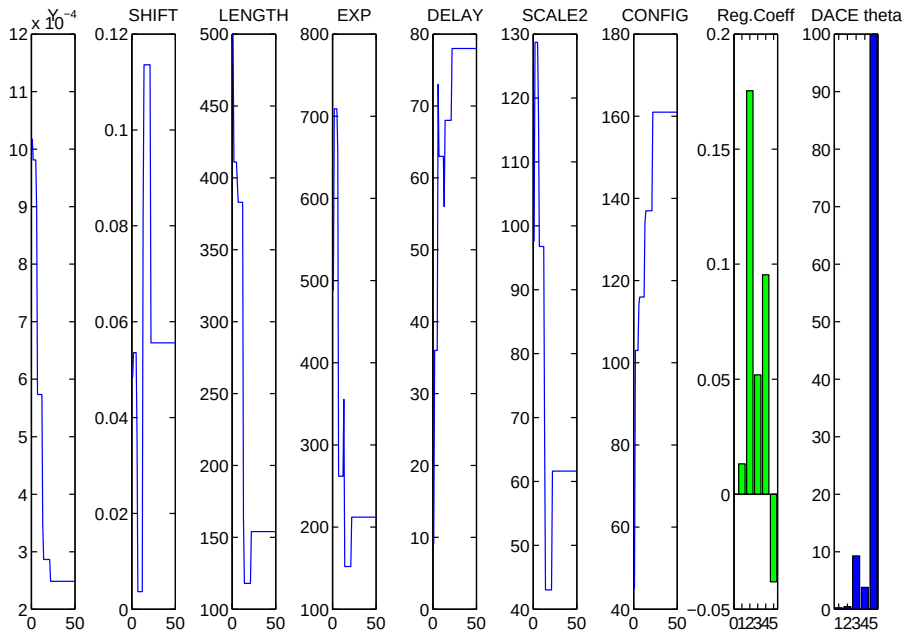

Bild 3: Sequentielle Parameter Optimierung. Während eines Laufs können die folgenden Informationen grafisch dargestellt werden (Abbildungen von links nach rechts): 1. Gütefunktion: Wie verändert sich die Vorhersagegüte? 2.-6. Parametereinstellungen für Verdunstung (SHIFT), Anz. Gewichte (LENGTH), Abklingfaktor (EXP), Verzögerung (DELAY) und Skalierung (SCALE2). 7. Nummer der besten Konfiguration (CONFIG). Die Parametereinstellungen werden durchnummeriert. Neue Einstellungen erhalten eine höhere Konfigurationsnummer. Steigen die CONFIG-Werte an, so werden neue Einstellungen erzeugt, die besser sind als die bisher gefundenen. 8. Koeffizienten eines linearen Regressionsmodells (üblicherweise β_i). Diese können als Indikatoren für die Wichtigkeit einzelner Faktoren herangezogen werden. 9. Koeffizienten eines stochastischen Prozessmodells (üblicherweise θ_i). Diese können ebenfalls als Indikatoren für die Wichtigkeit einzelner Faktoren herangezogen werden.

exemplarisch dargestellt. Das erste Schaubild zeigt die Veränderung der Gütefunktion, die nächsten Abbildungen zeigen die Veränderungen der fünf Parameter (SHIFT, LENGTH, EXP, DELAY und SCALE) während 50 Schritten der SPO. Da SPO sequentiell vorgeht und in jedem Schritt neue Parametereinstellungen generiert werden, kann das „Innovationspotential“ der SPO im siebten Schaubild (CONFIG) abgelesen werden. Die beiden letzten Abbildungen zeigen den Effekt der fünf Faktoren, einerseits gemessen als Regressionskoeffizienten eines klassischen linearen Regressionsmodells (Reg.Coeff) und andererseits als Korrelationskoeffizienten eines stochastischen Prozessmodells (DACE theta) [7]. Der Einfluss einzelner Faktoren kann auch durch einen Regresssionsbaum dargestellt werden [8]. Aus Abbildung 4 wird ersichtlich, dass der Faktor SCALE2, also die abschließende Skalierung, den größten Einfluss auf die Vorhersagegüte besitzt.

Dies stimmt überein mit den Werten der Korrelationskoeffizienten des stochastischen Prozessmodells. Die Koeffizienten des linearen Regressionsmodells legen den Schluss nahe, dass die Anzahl der Gewichte (LENGTH) den größten Einfluss besitzt. Dies ist nicht sonderlich überraschend. Aus anderen Studien ist uns bekannt, dass der Effekt der einzelnen Faktoren von dem zugrundeliegenden Modell abhängt und dass insbesondere Unterschiede zwischen linearen Regressionsmodellen und stochastischen Prozessmodellen

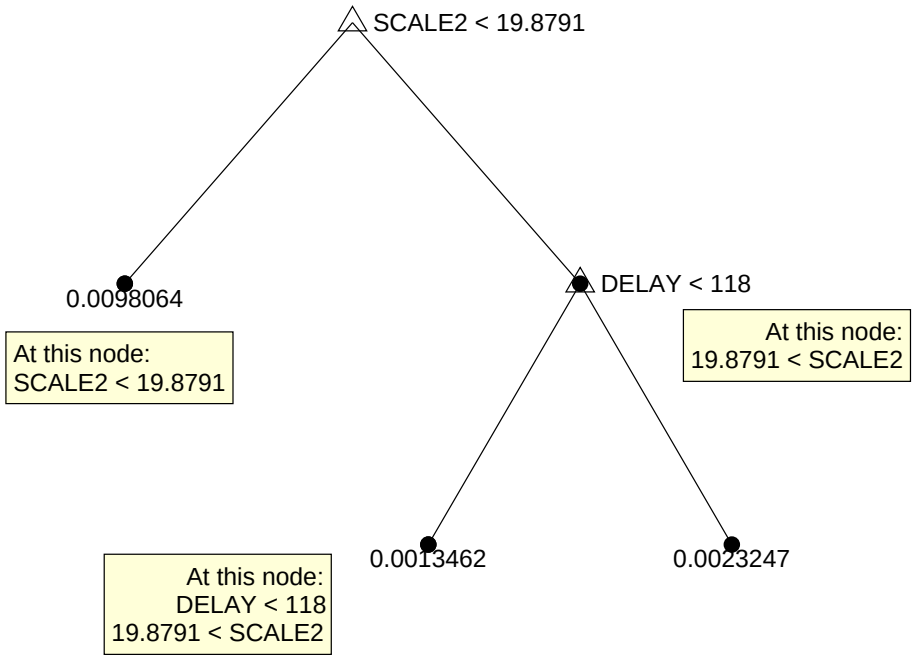

Bild 4: Regressionbaum zur Darstellung der Effekte der Filterparameter auf die Vorhersagegüte. Die besten Vorhersagen werden erzielt, wenn die Verzögerung (DELAY) kleiner als 118 und der Faktor für die Skalierung (SCALE2) größer als 19 gewählt werden.

auftreten.

3 Zusammenfassung

Als erstes Ergebnis erhalten wir schon eine relativ gute Prognose der mit SWMM simulierten Füllstände (Abbildung 2). Da die eingesetzten Verfahren - in diesem Fall neuronale Netze und FIR-Filter - mehrere Parametern benötigten, setzten wir Methoden der statistischen Versuchsplanung zu deren Bestimmung ein. Bei den NN konnten auch durch den Einsatz von SPO keine geeigneten Vorhersagemodelle generiert werden. Bei den FIR-Filter Modellen führte SPO schnell zu guten Parametereinstellungen und zu einem an das Problem angepassten Vorhersagemodell.

Der Anwender erhält somit statistisch abgesicherte Informationen, dass das empfohlene Prognosemodell besser geeignet ist als die anderen betrachteten Modelle und zusätzlich die Sicherheit, dass diese Aussagen nicht von zufälligen Messdaten abhängig sind, sondern dass diese Ergebnisse für ein großes Spektrum an unterschiedlichen problemspezifischen Parametern (wie z.B. die Bodenbeschaffenheit, Niederschlagsmengen) Gültigkeit behält.

Zukünftige Arbeiten untersuchen die Einsatzmöglichkeiten dieses Ansatzes für Daten, die nicht durch Simulationsmodelle wie SWMM erzeugt worden sind, sondern für empirisch gewonnene Messwerte. Hier treten durch Messungenauigkeiten und Messfehler weitere Schwierigkeiten auf, die momentan noch durch kein Prognosemodell zufriedenstellend gelöst werden können.

Danksagung

Für viele interessante Diskussionen und die Bereitstellung anwendungsbezogener Daten danken wir Dipl.-Ing. Tanja Hilmer und Dipl.-Ing. Andreas Stockmann.

Literatur

[1] Bongards, M.: Online-Konzentrationsmessung in Kanalnetzen – Technik und Betriebsergebnisse. Techn. Ber., Cologne University of Applied Sciences. 2007.

[2] U.S. Environmental Protection Agency: Storm Water Management Model. `http://www.epa.gov/ednnrmrl/models/swmm/index.htm`, Online; Stand 26.08.07. 2007.

[3] Jaeger, H.; Haas, H.: Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science* (2004), S. 78–80.

[4] Kleijnen, J. P. C.: *Statistical Tools for Simulation Practitioners*. New York NY: Marcel Dekker. 1987.

[5] Bartz-Beielstein, T.: *Experimental Research in Evolutionary Computation—The New Experimentalism*. Berlin, Heidelberg, New York: Springer. 2006.

[6] Bartz-Beielstein, T.; Preuss, M.: Moderne Methoden zur experimentellen Analyse evolutionärer Verfahren. In: *Proc. 16th Workshop Computational Intelligence* (Mikut, R.; Reischl, M., Hg.), S. 25–32. Universitätsverlag, Karlsruhe. 2006.

[7] Santner, T. J.; Williams, B. J.; Notz, W. I.: *The Design and Analysis of Computer Experiments*. Berlin, Heidelberg, New York: Springer. 2003.

[8] Breiman, L.; Friedman, J. H.; Olshen, R. A.; Stone, C. J.: *Classification and Regression Trees*. Monterey CA: Wadsworth. 1984.