Schätzmethode Verrechnungssteuer

Fassung vom 16. Februar 2018

Die Einnahmen aus der Verrechnungssteuer verzeichnen grosse Schwankungen und sind schwer prognostizierbar. Strukturelle Modelle oder das Abstellen auf einen langfristigen Durchschnittswert vermögen die Entwicklung der Einnahmen nicht abzubilden und noch weniger zu prognostizieren. Seit dem Voranschlag 2012 wird die Budgetierung der Verrechnungssteuereinnahmen mittels eines Zeitreihenmodells (robuste exponentielle Glättung) vorgenommen. Dieses Modell berücksichtigt bei der Erstellung der Prognose nur Vergangenheitswerte, wobei jüngere Werte ein grösseres Gewicht erhalten als ältere. Der Einfluss grosser Abweichungen („Ausreisser") wird reduziert, was zu einer Verstetigung der Budgetwerte führt. Zudem wird dem beobachtbaren Aufwärtstrend bei den Verrechnungssteuereinnahmen Rechnung getragen. Beträchtliche jährliche Abweichungen zwischen Budget und Rechnung können weiterhin auftreten, sollten sich aber im Zeitverlauf aufheben.

Inhaltsverzeichnis

1 Schwer prognostizierbare Verrechnungssteuereinnahmen

Die Einnahmen aus der Verrechnungssteuer verzeichnen grosse Schwankungen und sind schwer prognostizierbar.[1] In der Vergangenheit wurden die Einnahmen aus der Verrechnungssteuer mit verschiedenen Schätzmethoden budgetiert. Die Schätzungen erwiesen sich oft als klar zu tief und die Abweichungen zwischen Budget und Rechnung waren jeweils (in beiden Richtungen) gross. In den 90er Jahren, bis zur Einführung des Neuen Rechnungsmodells des Bundes (NRM), wiesen die Einnahmen zudem ein Zickzack-Muster auf: auf ein Jahr besonders hoher Einnahmen folgte unweigerlich ein Jahr deutlich tieferer Einnahmen. Dies wurde dadurch mitverursacht, dass per Ende Jahr keine Abgrenzungen vorgenommen wurden: hohe Eingänge waren bereits verbucht worden, die dazu gehörenden Rückerstattungsanträge liessen aber teilweise über ultimo auf sich warten. Dieses Problem wurde mit dem Übergang zum Forderungsprinzip grösstenteils behoben (nachdem die sogenannte Villiger-Regel[2] kurzfristig für eine gewisse Verstetigung gesorgt hatte). Die Schätzung gestaltete sich dennoch weiterhin schwierig. Nachdem sowohl strukturelle Modelle als auch eine Budgetierung gemäss langfristigem Durchschnitt[3] nicht erfolgreich waren,[4] wurde mit dem Voranschlag 2012 eine robuste Zeitreihenmethode eingeführt, welche im Folgenden beschrieben wird.

2 Anforderungen an die Schätzmethode

Die Schätzmethode sollte die folgenden Anforderungen erfüllen:

2.1 Unverzerrtheit

Die Schätzung sollte unverzerrt sein, das heisst, Über- und Unterschätzungen sollten sich im Zeitverlauf aufwiegen.

2.2 Prognosegenauigkeit

Es gilt zudem die Anforderung, dass die Schätzungen möglichst wenig um den tatsächlichen Wert streuen sollten. Diese Anforderung wird in der Regel operationalisiert durch den root-mean-square error (RMSE, den mittleren quadratischen Fehler), eine Verlustfunktion. Generell modelliert eine Verlustfunktion den „Schaden“, welcher durch eine Abweichung der Schätzung vom tatsächlichen Wert entsteht.

[1] Die Verrechnungssteuer macht ihrem Namen in diesem Sinne alle Ehre.

[2] Die Villiger-Regel (Dezember-Korrektur), benannt nach Alt-Bundesrat Kaspar Villiger, wurde per Ende 1999 eingeführt. Sie besagt, dass bestimmte, im Dezember eingereichte Deklarationen, also eigentlich fällige Forderungen, erst im Januar des Folgejahres verbucht werden, wenn die dazugehörenden Rückerstattungsforderungen erwartungsgemäss ebenfalls eintreffen werden. Ziel ist eine periodengerechte Erfassung von Einnahmen und Ausgaben.

[3] Zwischen 2004 und 2011 wurde die Verrechnungssteuer mit dem 10-Jahresdurchschnitt budgetiert.

[4] Auch eine KOF-Studie in 2001 brachte keine Verbesserungsvorschläge.

2.3 Verstetigung

Aus Rücksicht auf die Budget- bzw. Ausgabenpolitik sollte die Prognose keine starken Schwankungen aufweisen. Der Einfluss von Ausnahmejahren mit sehr hohen / tiefen Einnahmen sollte deshalb begrenzt werden.

2.4 Anpassung an Entwicklungen

Die Schätzungen sollten neue Einnahmenentwicklungen berücksichtigen, damit die Prognosen diesen Einnahmenentwicklungen zeitlich nicht zu weit hinterher hinken. Eine gewisse zeitliche Verzögerung ist schon deshalb unvermeidbar, da die Budgetprognose bereits im Juni, in Unkenntnis der Einnahmen des laufenden Jahres, für das nächste Jahr aufgestellt wird.

Insgesamt wachsen die Steuereinnahmen des Bundes, bereinigt um Sonderfaktoren und Massnahmen, etwa im Schritt mit dem BIP. Auch die der Verrechnungssteuer unterliegenden Zinsen und Gewinnausschüttungen nehmen tendenziell zu, wenn auch nicht unbedingt im Gleichschritt mit dem BIP. Deshalb ist im Normalfall ein Wachstum der Einnahmen über den Finanzplanhorizont zu erwarten. Eine Budgetierung mittels fixem Wert scheidet deshalb aus.

3 Alternative Ansätze zur Schätzung der Verrechnungssteuereinnahmen

Grundsätzlich stehen drei verschiedene Methoden zur Auswahl, um die Einnahmen aus der Verrechnungssteuer zu modellieren und anschliessend zu prognostizieren:

3.1 Strukturelle Modelle

In einem strukturellen Modell der Steuereinnahmen wird versucht, die Einnahmen durch gewisse korrelierte Zeitreihen (z.B. BIP, Börsenumsätze, Einkommen, Zinsen, Gewinnausschüttungen) zu erklären und anschliessend zu schätzen. Die Prognosen der erklärenden Faktoren bestimmen demzufolge die Prognose der Steuereinnahmen.

In der Vergangenheit sind sämtliche Versuche, die Einnahmen aus der Verrechnungssteuer mit einem strukturellen Modell zu erfassen und prognostizieren, fehlgeschlagen: der Ertrag aus der Verrechnungssteuer korreliert nicht genügend mit den verfügbaren Wirtschaftsdaten. Aus diesem Grund wird diese Alternative in der vorliegenden Notiz nicht weiter ausgeführt.

3.2 Statisches Modell

Ein statisches Modell geht davon aus, dass sich die interessierende Grösse nicht (dauerhaft) verändert. Allfällige Abweichungen vom langfristigen Wert sind demzufolge zufällig und vorübergehend. Der zwischen 2004 und 2011 verwendete fixe Budgetwert von 3 Mrd. entspricht diesem Modelltyp. Gewählt wurde dieses Modell, weil man zur Einsicht gelangt war, dass strukturelle Modelle nicht geeignet waren, um den historischen Verlauf der Verrechnungssteuereinnahmen zu erklären. Zudem nahm man an, dass sich bei Verwendung des langfristigen Mittels als Budgetwert Über- und Unterschätzungen in Zukunft abwechseln und kompensieren würden.

Allfällige, durch steuerpolitische Massnahmen verursachte, Strukturbrüche (bzw. Niveauverschiebungen) müssten in diesem Modell ad hoc (durch Ergänzung um einen Massnahmen-Faktor) berücksichtigt werden. In der Periode 2004-2011 wurden keine solchen ad hoc Anpassungen vorgenommen.

3.3 Zeitreihenmodelle

In einem Zeitreihenmodell wird eine Zeitreihe (wie Steuereinnahmen) mit vergangenen Werten derselben Zeitreihe erklärt und prognostiziert. Die Zukunft wird hier sozusagen direkt aus der Vergangenheit extrapoliert. Allfällige Strukturbrüche durch steuerpolitische Massnahmen müssten ad hoc berücksichtigt werden. Je nach Modell und je nach „Gedächtnis" des Zeitreihenmodells passt sich die Prognose mehr oder weniger schnell den veränderten Bedingungen an.

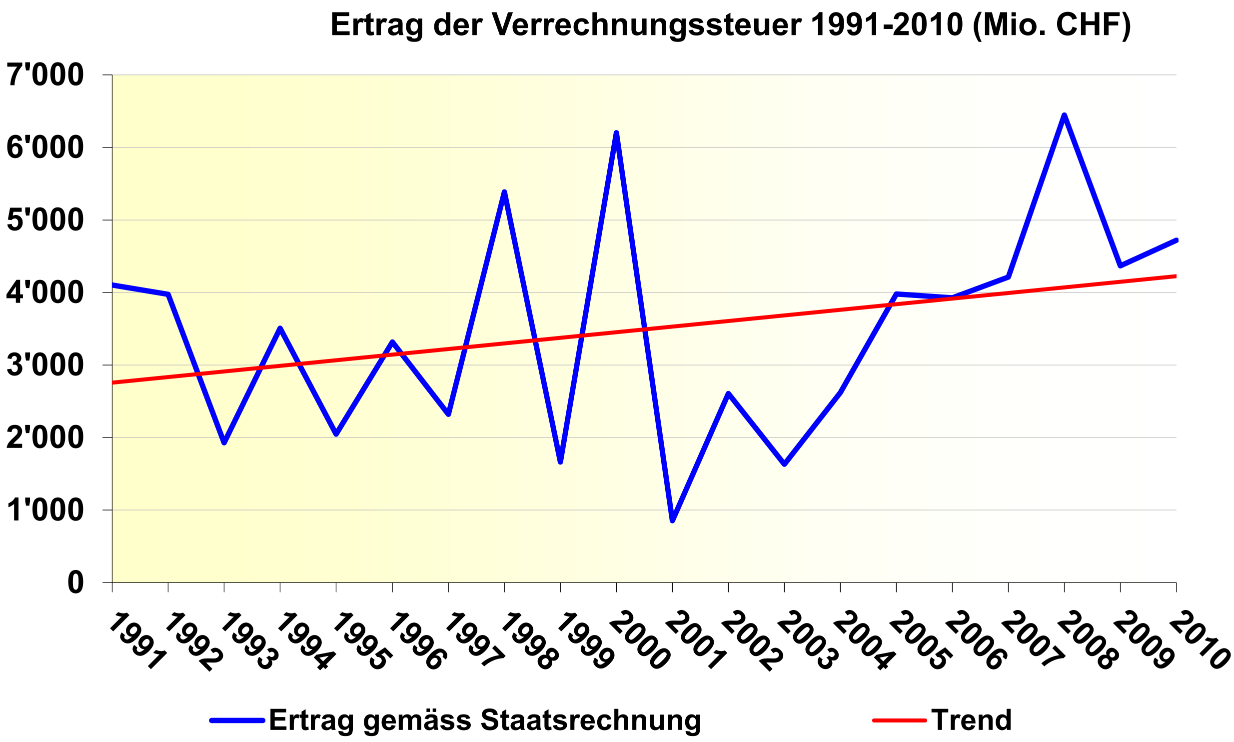
Die Ableitung von Aussagen zur Zukunft aus reinen Vergangenheitswerten führt prinzipiell dazu, dass die so gewonnenen Prognosen der Entwicklung immer hinterher hinken. Bei der Wahl des konkreten Zeitreihenmodells ist darauf zu achten, dass dieser Lag möglichst gering bleibt, ohne jedoch die Stabilität der Prognosen zu gefährden. Nur strukturelle Modelle schauen einigermassen vorwärts, allerdings brauchen diese ihrerseits Prognosen für die bestimmenden Grössen, welche zwangsläufig fehlerbehaftet sein werden.

4 Verwendetes Schätzverfahren: Robuste Holt-Winters Methode (RHW)

4.1 Allgemeine Eigenschaften der Methode

Im Modell der *exponentiellen Glättung* werden die jüngsten Beobachtungen stärker gewichtet, wodurch das „Gedächtnis" der Prognose je nach Ausgestaltung tendenziell reduziert wird. Bei weit zurückliegenden Werten reduziert sich das Gewicht der Beobachtungen exponentiell. Die reine Methode der exponentiellen Glättung funktioniert aber nur gut, wenn die Zeitreihe grundsätzlich stationär ist, das heisst, es darf kein Trend vorhanden sein. In der unten stehenden Grafik ist der Ertrag aus der Verrechnungssteuer für die 20 Jahre von 1991-2010 dargestellt, als die neue Methode entwickelt wurde. Es ist klar ersichtlich, dass ein leichter Aufwärtstrend vorhanden ist.

Ein eventuell in den Daten vorhandener *linearer Trend* wird durch die so genannte Holt-Methode, eine flexiblere Variante der exponentiellen Glättung, berücksichtigt. Eine Analyse der Einnahmenentwicklung in dieser Zeitperiode zeigt, dass trotz der grossen jährlichen Schwankungen der unterliegende Trend die Schätzung massgeblich beeinflusst. Der letzte Rechnungswert wirkt sich in diesem Modell indirekt über den Schätzfehler auf die Schätzung für die nächsten Perioden aus.



Viele Zeitreihen verzeichnen nicht nur einen Trend, sondern weisen zudem auch Extremwerte bzw. Ausreisser auf. Im Standard-Modell der linearen exponentiellen Glättung nach Holt (wie in Standard-Regressions- und Durchschnittsmodellen) können (vor allem jüngere) Ausreisser die Prognose destabilisieren und die Genauigkeit reduzieren. Dies würde dem Ziel einer Verstetigung der Budgetierung entgegenlaufen. Eine Weiterentwicklung des Holt-Modells macht die Schätzungen *robust*: Dazu werden in einem ersten Schritt Ausreisser begrenzt, indem anhand der empirisch ermittelten Varianz für die nächste Beobachtung eine Ober- und Untergrenze angesetzt wird. In einem zweiten Schritt wird die Verlustfunktion, welche zur Bestimmung der optimalen Parameter der Methode minimiert wird, weniger ausreisserbetont spezifiziert. Da die Jahresergebnisse der Verrechnungssteuer grössere Ausschläge verzeichnen, kann nur ein robustes Verfahren deren verzerrende Wirkung dämpfen. Diese Methode wird als robuste Holt-Winters Methode (nach Holt und Winters 1959, RHW) bezeichnet. Sie wurde von Gelper, Fried und Croux (2010) entwickelt.

4.2 Die Methode im Detail

Die Beschreibung in diesem Abschnitt folgt Gelper, Fried und Croux (2010). Exponentielle Glättung kann durch folgende Gleichung dargestellt werden:

$$\tilde{y}_t = \lambda y_t + (1-\lambda)\tilde{y}_{t-1}$$

Das heisst: der geglättete Wert $\tilde{y}_t$ für eine Reihe zum Zeitpunkt *t* ist eine gewichtete Kombination vom aktuellen Wert und dem vorangegangenen geglätteten Wert. λ ist der Glättungsparameter, welcher zwischen null und eins variieren kann. Wenn λ gleich null ist, wird der vergangene Wert eingesetzt: die Glättung ist vollständig. Wenn λ gleich eins ist, wird der aktuelle Wert genommen: es erfolgt keine Glättung.

Durch wiederholtes Ersetzen des jeweils vorangegangenen geglätteten Wertes mit dem um eine Periode verzögerten Wert der Gleichung resultiert (falls $\tilde{y}_1 = y_1$)

$$\tilde{y}_t = \sum_{j=0}^{t-2} \lambda(1-\lambda)^j y_{t-j} + (1-\lambda)^{t-1} y_1$$

Hieraus ist ersichtlich, dass der geglättete Wert zum Zeitpunkt *t* mit exponentiell abfallenden Gewichten von allen vorangegangenen beobachteten Werten abhängt.

Eine Zeitreihe mit Trend lässt sich durch folgende zwei Gleichungen repräsentieren, wobei es nun zusätzlich eine lokale Trendvariable F_t gibt:

$$\tilde{y}_t = \lambda_1 y_t + (1-\lambda_1)(\tilde{y}_{t-1} + F_{t-1})$$

$$F_t = \lambda_2(\tilde{y}_t - \tilde{y}_{t-1}) + (1-\lambda_2)F_{t-1}$$

λ_1 und λ_2 sind die beiden Glättungsparameter; sie bestimmen das relative Gewicht der früheren und späteren Beobachtungen. Bei $\lambda = 1$ fällt das Gewicht voll auf die letzte Beobachtung, d.h. die Geschichte spielt keine Rolle. Bei $\lambda = 0$ hingegen zählt jeweils nur der vorangehende Wert, also effektiv nur der Anfangswert: die Entwicklung der Zeitreihe spielt keine Rolle.

Die Prognose $\hat{y}_{t+h\ |t}$ für die Periode *t+h* zum Zeitpunkt *t* wird wie folgt ermittelt:

$$\hat{y}_{t+h\ |t} = \tilde{y}_t + hF_t$$

Das System ist rekursiv lösbar. Die Holt-Winters Methode ist jedoch anfällig gegenüber Ausreissern: eine sehr hohe oder sehr tiefe Beobachtung verzerrt die geglättete Zeitreihe nach oben bzw. nach unten und reduziert in der Folge die Prognosequalität. Die sogenannte robuste Holt-Winters Methode (RHW) berücksichtigt die Existenz von Ausreissern und reduziert ihren Einfluss auf die geglättete Reihe und somit auf die Prognosen. Die Idee dabei ist, dass die Zeitreihe zuerst bereinigt wird, bevor die exponentielle Glättung stattfindet. Die Gleichungen für die RHW sind dementsprechend:

$$\tilde{y}_t = \lambda_1 y^*_t + (1-\lambda_1)(\tilde{y}_{t-1} + F_{t-1})$$

und

$$F_t = \lambda_2(\tilde{y}_t - \tilde{y}_{t-1}) + (1-\lambda_2)F_{t-1}$$

wobei y_t^* der bereinigte Wert zum Zeitpunkt *t* darstellt.

Die bereinigte Zeitreihe y_t^* ergibt sich aus der folgenden Gleichung:

$$y_t^* = \psi\left(\frac{y_t - \hat{y}_{t\ |t-1}}{\hat{\sigma}_t}\right)\hat{\sigma}_t + \hat{y}_{t\ |t-1}$$

wobei die ψ-Funktion auf die standardisierten Prognosefehler $\frac{y_t - \hat{y}_{t\ |t-1}}{\hat{\sigma}_t}$ für die jeweils nächste Periode zur Anwendung kommt. Die Prognosen werden wie vorher beschrieben ermittelt. Die Grössenordnung der Prognosefehler wird durch $\hat{\sigma}_t$ geschätzt. Die ψ-Funktion reduziert den Einfluss von Ausreissern. Es wird die Huber ψ-Funktion verwendet (wobei $sign(x)$ das

Vorzeichen von *x* ergibt):

$$\psi(x) = \begin{cases} x & \text{wenn} |x| < k \\ sign(x)k & \text{sonst} \end{cases}$$

Diese Funktion ersetzt unerwartet hohe oder tiefe Werte mit einem wahrscheinlicheren Wert. Wird für die Konstante k den Wert zwei gewählt, entspricht dies implizit einer angenommenen Normalverteilung der Prognosefehler für die jeweils nächste Periode (one-step ahead forecast error) $r_t = y_t - \hat{y}_{t\ |t-1}$. In anderen Worten, unter der Annahme eines normalverteilten „one-step-ahead forecast errors" impliziert $k = 2$, dass eine Beobachtung ausserhalb des 95%-Konfidenzintervalls der Prognose als Ausreisser angesehen wird. Der standardisierte Prognosefehler $\frac{y_t - \hat{y}_{t\ |t-1}}{\hat{\sigma}_t}$ wird dann mit einem von k abhängigen Wert ersetzt; bei $k = 2$ also mit einem Wert, der zwei Standardabweichungen vom prognostizierten Wert $\hat{y}_{t|t-1}$ entfernt liegt. Wenn k unendlich gross ist, entspricht die robuste Methode der klassischen Methode, d.h. keine Korrektur von Ausreissern. Die geschätzte Grössenordnung der Prognosefehler $\hat{\sigma}_t^2$ wird wie folgt berechnet, wobei die Grössenordnung im Zeitverlauf leicht variieren kann:

$$\hat{\sigma}_t^2 = \lambda_\sigma \rho\left(\frac{r_t}{\hat{\sigma}_{t-1}}\right)\hat{\sigma}_{t-1}^2 + (1-\lambda_\sigma)\hat{\sigma}_{t-1}^2$$

Hier wird von einer beschränkten Verlustfunktion ρ Gebrauch gemacht. Eine sogenannte Bi-Weight ρ -Funktion wird gewählt:

$$\rho(x) = \begin{cases} c_k\left(1-\left(1-(x/k)^2\right)^3\right) & \text{if } |x| \le k \\ c_k & \text{sonst} \end{cases}$$

Hierbei ist c_k eine Konstante, welche Konsistenz des Grössenordnungsparameters für normalverteilte Fehler sicherstellt. Bei der üblichen Wahl von $k = 2$ folgt $c_k = 2.52$. Der Schätzer für die Grössenordnung entspricht nun einem τ^2-Schätzer wie in Yohai und Zamar (1988), rekursiv berechnet. Dieser τ^2-Grössenordnungsschätzer (scale estimator) ist sehr robust und gleichzeitig fast gleich effizient wie die Standardabweichung.

Wenn nun die Bausteine kombiniert werden, kann das rekursive System durch zwei Gleichungen zusammengefasst werden:

$$\widetilde{y}_t = \lambda_1 \psi\left(\frac{y_t - (\hat{y}_{t-1} + F_{t-1})}{\hat{\sigma}_t}\right)\hat{\sigma}_t + \widetilde{y}_{t-1} + F_{t-1}$$

$$F_t = \lambda_2(\widetilde{y}_t - \widetilde{y}_{t-1}) + (1-\lambda_2)F_{t-1}$$

Es sei darauf hingewiesen, dass die gewählten Funktionen für ψ, ρ sowie der Grössenordnungsschätzer $\hat{\sigma}_t$ in der Literatur zur robusten Statistik Standard sind.

Schliesslich müssen nicht nur die Aktualisierungsgleichungen robust sein, die Anfangswerte sollten dies auch sein. Für die Holt-Winters Glättung wird eine wiederholte Medianschätzung (siehe Siegel (1982)) eingesetzt. Die Anfangswerte für den geglätteten Niveauwert $\widetilde{y}_m$ und den

Trend F_m betragen:

$$\tilde{y}_m = \hat{\alpha}_0 + \hat{\beta}_0 m$$

$$F_m = \hat{\beta}_0$$

wobei

$$\hat{\beta}_0 = \mathrm{med}_{i \neq j} \frac{y_i - y_j}{i - j}$$

$$\hat{\alpha}_0 = \mathrm{med}_i (y_i - \hat{\beta}_0 i)$$

für *i,j* = 1,…,*m.* Die wiederholte Median-Regression (repeated median regression) schneidet bei der Glättung von Zeitreihen gut ab.

Der Anfangswert für die rekursive Grössenordnungsschätzung ($\hat{\sigma}_t^2$) kann durch den Median der absoluten Abweichungen (median absolute deviation (MAD)) der Regressionsrestgrössen in der Anlaufperiode ermittelt werden.

4.3 Spezifikation des Schätzmodells

Für die Schätzungen wird ein spezifisches Programm, geschrieben in der statistischen Software R, verwendet. Dieses Programm wurde der ESTV von Prof. S. Gelper zur Verfügung gestellt.

Für die Schätzungen werden folgende Einstellungen verwendet:

- Die beiden Glättungsparameter λ_1 und λ_2 durften anfänglich zwischen 0.1 und 0.3 variieren. Eine solche Einschränkung gilt als üblich; damit wird eine sprunghafte, zufällige Änderung der λ von einem Jahr zum nächsten vermieden. Die Optimierung findet auf der Stufe einer Dezimalstelle statt. Zwei Dezimalstellen haben in der Simulation nur unwesentliche Unterschiede hervorgebracht und die Literatur stellt i.A. auf eine Dezimalstelle ab. Nach den ersten paar Jahren der Anwendung wurde das optimale Modell mit den Anpassungsparametern λ_1=0.3 und λ_2=0.2 festgelegt. Diese Werte wurden deshalb fixiert, weil eine jährliche Neuschätzung zu einer unerwünschten zusätzlichen Varianz des Niveauwerts sowie des Trends führen könnte, da eine alternative Kombination der beiden Werte die um eine Beobachtung erweiterte Zeitreihe als statistisch optimal gelten würde. Die alternative Kombination wäre indes nur leicht überlegen. Die resultierende Änderung des Trendwerts würde von Jahr zu Jahr zu starken Schwankungen der Finanzplanwerte führen, was dem Ziel der Verstetigung zuwiderlaufen würde.

- Für die RHW-Periode (worauf die robuste exponentielle Glättung basiert) werden die letzten verfügbaren Jahre herangezogen. Die Anlaufperiode (welche die Anfangswerte bestimmt) entspricht den ersten 20 Jahren 1950-1969, bis zum Anfang der RHW-Periode. Damit wird die Ausgangslage nicht jährlich geändert, was sonst zu einer künstlich verursachten Varianz in den Parametern führen könnte.

- Die Prognose besteht aus dem geglätteten Wert für das letzte verfügbare Jahr plus zweimal den letzten Trendwert, da der Voranschlag für das nächste Jahr immer bereits festgelegt werden muss, bevor der Rechnungswert des laufenden Jahres bekannt ist. Eine allfällige manuelle Hochrechnung für das laufende Jahr wird somit nicht berücksichtigt.

- Es findet auch keine Berücksichtigung von steuerpolitischen Massnahmen statt. Die Auswirkungen von Massnahmen fliessen durch die Schätzmethode im Zeitverlauf in die Prognosen ein.

5 Literaturverzeichnis

Gelper S., Fried R., und Croux C., 2010. Robust Forecasting with Exponential and Holt-Winters Smoothing, Journal of Forecasting 29, S. 285-300, 2010.

Holt C., 1959. Forecasting seasonals and trends by exponentially weighted moving averages. ONR Research Memorandum 52.

Siegel A., 1982. Robust regression using repeated medians. Biometrika 68, S. 242-244.

Winters P., 1960. Forecasting sales by exponentially weighted moving averages. Management Science 6, S. 324-342.

Yohai V., Zamar R., 1988. High breakdown estimates of regression by means of the minimization of an efficient scale. Journal of the American Statistical Association 83, S. 406-413.