Statistische Verfahren:
Projekt 4 - Nahinfrarotspektroskopie I

Tobias Giesemann
Ferdinand Rewicki
Moritz Preuß

March 30, 2019

Abstract

In der vorliegenden Arbeit wird eine Methode zur Ableitung einer Kalibierfunktion zur Messung des Stickstoffgehaltes in Bodenproben vorgestellt. Basierend auf den zugrunde liegenden physikalischen und chemischen Eigenschaften der Nahinfrarotspektroskopie wurde ein lineares Modell formuliert dessen Prediktoren mit Hilfe von Mallow's Cp ausgewählt wurden. ....

Contents

Statistische Verfahren:
Projekt 4.1 - Nahinfrarotspektroskopie I

Tobias Giesemann
tobias.giesemann@uni-jena.de

Ferdinand Rewicki
...@gmail.com

Moritz Preuß
moleo.preuss@gmail.com

Abstract

In der vorliegenden Arbeit wird eine Methode zur Ableitung einer Kalibierfunktion zur Messung des Stickstoffgehaltes in Bodenproben vorgestellt. Basierend auf den zugrunde liegenden physikalischen und chemischen Eigenschaften der Nahinfrarotspektroskopie wurde ein lineares Modell formuliert dessen Prediktoren mit Hilfe von Mallow's Cp ausgewählt wurden. ....

1 Einleitung

Die Zusammensetzung des Boden ist ein wesentlicher Faktor für einen ertragreichen und nachhaltigen Anbau von Pflanzen in der Landwirtschaft. Wichtige Parameter hierfür sind die Anteile des im Boden gebundenen Stickstoffs (N) und organische Kohlenstoffs (SOC), durch deren Messung Erkenntnisse über die Fruchtbarkeit des Bodens und die Auswirkungen der Bodennutzung gewonnen werden können.[PDD+13] Die Messung der genannten Parameter ist durch sogenannte Fraktionierung von Bodenprobe möglich. Etablierte Messverfahren sind jedoch kostenintensiv, nicht vollständig standardisiert, und weisen eine schlechte Reproduzierbarkeit in verschiedenen Laboren auf.[PDD+13] Aus diesem Grund sind Messverfahren notwendig, welche eine zuverlässige und effiziente Ermittlung der Menge des organischen Kohlenstoffs und gebundenen Stickstoffs im Bodens zulassen. Ein seit den sechziger Jahren vielfach eingesetztes Messverfahren ist die sogenannten Nahinfrarotspektroskopie. [AHJ10] Dieses erlaubt die Schätzung der aufwendig bestimmbaren Parameter mit Hilfe der leicht messbaren Nahinfrarotspektren. Die Bestimmung und Validierung eines hierfür notwendigen Modells für den Stickstoffgehalt in der Bodenprobe sind Gegenstand dieser Arbeit.

2 Hintergrund

2.1 Gebundener Stickstoff

In der Natur geschieht eine fortwährender Austausch von Stickstoff zwischen Lebewesen, Boden und Atmosphäre. Ein Großteil des vorhandenen Stickstoffs liegt in gebundener Form im Erdboden vor und ist ein unentbehrlicher Nährstoff für Pflanzen und Lebewesen. Er wird von Pflanzen beim Wachstum aus der Erde aufgenommen und beim Absterben wieder freigesetzt. Zur Steigerung der Ernte ist eine Anreicherung der Bodens mit Nitrat durch der Einsatz von Düngemitteln daher gängige Praxis in der Landwirtschaft.[Umw17] Dies kann dann zu Problemen führen, wenn zu einer Übersättigung des Bodens mit Stickstoff kommt. Durch Ausschwemmen des in Form von Nitrat (NO_3^-) und Ammonium (NH_4^+) gebundenen Stickstoffs kann dieses ins Grundwasser gelangen und eine Gefahr für die Umwelt darstellt. Sowohl für den Erfolg der Landwirtschaft als auch für den Schutz der Umwelt ist daher eine zuverlässige und effiziente Ermittlung des Nitratgehalts im Bodens von entscheidender Bedeutung. Der Messung des Stickstoffs liegen folgende chemische Zusammenhänge zu Grunde:

Die Stoffmengenkonzentration von Stickstoff $c_{(N)}$ lässt sich berechnen durch,

$$c^{(N)} \coloneqq \frac{n^{(N)}}{V}$$

wobei V das Volumen der Lösung und $n_{(N)}$ die enthaltenen Stoffmenge von Stickstoff ist.

Für eine gegebene Stoffmengenkonzentration c_0 und Stoffmenge n_o einer Probe, definieren wir den Stoffmengenanteil $y_{(N)}$ von Stickstoff als,

$$y_{(N)} \coloneqq \frac{c^{(N)}}{c_0} = \frac{n^{(N)}}{n_0}$$

2.2 Nahinfrarotspektroskopie

Bei der Nahinfrarotspektroskopie kommen elektromagnetische Wellen im Bereich zwischen 120 THz und 400 THz bzw. 2.500 nm und 750 nm zum Einsatz. [AHJ10] Das Messverfahren nutzt die Tatsache, dass bei der Bestrahlung einer Probe Teile des Lichts reflektiert, hindurch gelassen

oder absorbiert werden. Von besonderem Interesse ist hierbei die Reflexion des Lichts, welche sich in den zwei Komponenten Spiegelreflexion und diffuse Reflexion unterscheiden lässt. Aus den Teilen des Lichts welche diffus reflektiert werden, können aufgrund der größeren Eindringtiefe Informationen über die Beschaffenheit der Probe gewonnen werden. [AHJ10] Für eine Wellenlänge λ ist das relative Reflexionsvermögen $\delta(\lambda)$ definiert als:

$$\delta\colon (0,\infty) \to (0,\infty), \qquad \delta(\lambda) \coloneqq \frac{P_{\mathrm{r}}(\lambda)}{P_s}$$

wobei $P_{\mathrm{r}}(\lambda)$ die Reflexion einer Probe und P_0 die Reflexion eines Materials mit einem Reflexionsanteil nahe 100% ist.

Weiterhin lässt sich unter Anwendung des Beer'schen Gesetzes ein approximativer Zusammenhang zwischen dem relativen Reflexionsvermögen und der Stoffmenge $c_{(N)}$ des Stickstoffs herstellen. In eine Probe mit n verschiedenen Stoffen sei c_i die Stoffmengenkonzentration des $i$ten in der Probe enthaltenen Stoffes. Dann sei $\varepsilon_i(\lambda)$ ein Koeffizient mit $i \in \mathbb{N}, i \leq n$ sodass

$$-\log \delta(\lambda) = -\log \frac{P_{\mathrm{r}}(\lambda)}{P_s} = \sum_{i=1}^{n} \varepsilon_i(\lambda) c_i$$

3 Methodik

3.1 Datensatz

Der für die Modellwahl verwendete Datensatz beinhaltet die logarithmierten relativen Reflexionswerte $-\delta(\lambda)$ bei Wellenlängen zwischen 1400 nm und 2672 nm in einem Abstand von 4 nm sowie die Stoffmengeanteile $y^{(N)}$, $y^{(SOC)}$ und entsprechenden pH-Werte von insgesamt 533 Proben. Informationen über die chemische Zusammensetzung der Probe in den Reflexionswerten im Nahinfrarotbereich sind stark überlagert. [AHJ10] Aus diesem Grund ist eine sorgfältige Auswahl relevanter Wellenlängen von besonderer Bedeutung für die Erstellung eines zuverlässigen Modells.

3.2 Statistisches Modell

Sei $n \in \mathbb{N}$ die Größe des Datensatzes und $k \in \mathbb{N}$ mit $k < n$ die Anzahl der Wellenlängen im Datensatz. Entsprechend Abschnitt 3.1 definieren wir dann die Einflussgröße x_{ij} als für die $i$te Probe und $j$te Wellenlänge als

$$x_{ij} \coloneqq -\lg \delta_i(\lambda_j)$$

für jedes $i, j \in \mathbb{N}, i \leq n, j \leq k$.

Der Stoffmengenanteil des Stickstoffs $y^{(N)}$ stellt die Zielgröße unseres späteren Modells dar. Wie definieren hierfür den Vektor y_i den Stoffmengenanteil der $i$ten Probe als den n dimensionalen Vektor

$$y^{(N)} \coloneqq \left(y_i^{(\mathrm{N})}\right)$$

Nachdem wir sowohl die Einflussgrößen als auch die Zielgröße für das lineare Modell definiert haben lassen sich diese nun in Zusammenhang bringen. Es ist plausibel anzunehmen, die sich sie Zielgröße durch einen Linearkombination der Einflussgrößen beschreiben lässt. Hierfür definieren wir zunächst $Y^{(\mathrm{N})}$ als einen zufälligen Vektor von $y^{(\mathrm{N})}$

$$\mathbb{E}\, Y^{(\mathrm{N})} \coloneqq \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j$$

Zudem ist es notwendig eine Variable $\varepsilon^{(\mathrm{N})}$ einzuführen welchen den Zufall der Messungen beschreibt. In Matrixschreibweise lässt sich dies als die Designmatrix $\mathbb{X} \in \mathbb{R}^{n\times(k+1)}$, dem Parametervektor $\beta \in \mathbb{R}^{k+1}$ und dem stochastisch verteilten Parameter $\varepsilon^{(\mathrm{N})}$ darstellen, sodass gilt,

$$Y^{(\mathrm{N})} = \mathbb{X}\beta + \varepsilon^{(\mathrm{N})}$$

mit

$$\mathbb{E}\,\varepsilon^{(\mathrm{N})} = 0, \qquad \operatorname{cov}\varepsilon^{(\mathrm{N})} = (\sigma^2)^{(\mathrm{N})}\mathrm{I}$$

wobei $(\sigma^2)^{(\mathrm{N})} \in (0,\infty)$. Weiterhin soll angenommen werden, dass $\varepsilon^{(\mathrm{N})}$ normalverteilt ist mit

$$\varepsilon^{(\mathrm{N})} \sim \mathcal{N}\left(0, (\sigma^2)^{(\mathrm{N})}\mathrm{I}\right)$$

sodass sich für das Gesamtmodell gilt

$$Y^{(\mathrm{N})} \sim \mathcal{N}\left(\mathbb{X}\beta^{(\mathrm{N})}, (\sigma^2)^{(\mathrm{N})}\mathrm{I}\right)$$

3.3 Modellwahl im Falle der NIR-Spektroskopie

Seien $Y \in \mathbb{R}^{\mathrm{n}}$ die Zielgröße in einem statistischen Modellwahlverfahrens und $\mathbb{X} \in \mathbb{R}^{\mathrm{n}\times\mathrm{d}}$ eine Designmatrix. Zur Wahl einer geeigneten Menge von k Einflussparametern auf die Zielgröße y_i wird klassischerweise die Modellwahl über eine hierarchische Aufstellung von linearen Modellen erreicht. Beginnend mit dem minimalen Modell $E(Y_i) = \beta_0$ werden nach und nach neue potenzielle Einflussparameter x_{ik} hinzugefügt. Zu jedem dieser neuen x_{ik} wird dann eine Teststatistik aufgestellt, die darauf hinweist, ob der gewählte Parameter wichtig ist ist oder nicht. Dabei ist die Nullhypothese, dass x_{ik} keinen Einfluss auf die Zielgröße hat: $H_0 = \beta_k = 0$ und wird abgelehnt, falls $H_1 = \beta_k \neq 0$ zutrifft. Dies wird über die T-Teststatistik erreicht, wobei für den Fall, dass H_0 richtig ist, gilt:

$$\frac{\hat{\beta_k}}{\sqrt{\sigma^2(\mathbb{X}^T\mathbb{X})^{-1}_{kk}}} \sim t_{n-(k+1)}.$$

Dieses Verfahren ist vor allem dann besonders gut geeignet, wenn man bereits theoretisch fundierte Annahmen über die Einflussgrößen machen kann. Mit diesem Modellwahlverfahren ergeben sich hier allerdings einige Schwierigkeiten, wobei die für unseren Fall besonders schwerwiegenden herausgehoben werden: In dieser Arbeit haben wir es mit einer großen Anzahl potenzieller Einflussvariablen auf dem Nah-Infrarotspektrum zu tun. A priori kann schwer eine inhaltliche Deutung vorgenommen werden, die gewisse Wellenlängen bevorzugt. Daher ist eine hierarchische Modellwahl mit wenigen, theoretisch begründeten Einflussvariablen nicht möglich. Demnach muss in dieser Arbeit die Anzahl der möglichen Einflussgrößen stark erhöht werden und hier bekommen wir ein Problem mit der T-Teststatistik. Es ließen sich sehr viele unterschiedliche Kombinationen von Einflussgrößen aufstellen und in eine hierarchische Form bringen. Doch da wir bei der T-Teststatistik ein zufälliges Intervall konstruieren, gegen das unsere Hypothese getestet wird, wird die Wahrscheinlichkeit bei oft wiederholten Tests fälschlicherweise die Nullhypothese abzulehnen mit Anzahl der Versuchen immer größer. An ein automatisiertes Modellwahlverfahren, das in dieser Arbeit von Vorteil ist, ist also mittels des T-Tests nicht zu erreichen [Sch19]. Stattdessen bietet sich eine Modellwahl basierend auf dem erwarteten Prognosefehler ("sum of prediction squared error", SPSE) an:

$$SPSE := E(\sum_{i=1}^{n}(Y_{i+n} - x_i^{(M)}\hat{\beta}_i{}^{((M)})^2)$$

Hierbei sind die Werte in Y_{i+n} neue Beobachtungen zum Erwartungsvektor x_i und $x_i^{(M)}\hat{\beta}_i{}^{((M)}$ ist sind die Prognosewerte aus dem zu testenden Modell M. Der Prognosefehler lässt sich in 3 Terme zerlegen: Einen irreduzierbaren Prognosefehler, der unabhängig von dem momentan betrachteten Modell ist, einen Biasterm, der die Abweichung des aktuellen Modells M vom Prognosemodell als Summe der quadrierten Prognose-Verzerrungen anzeigt und einen Varianzterm, der die Ungenauigkeiten widerspiegelt, die sich aus der Schätzung von $p = (|M| + 1)$ unbekannten Parametern ergibt:

$$SPSE^{(M)} = n\sigma_{full}^2 + p\sigma_{full}^2 + (bias^{(M)})^2$$

Der SPSE lässt sich über unterschiedliche Wege schätzen:
(1)mithilfe neuer Beobachtungen,
(2) (wiederholter) Zerlegung der Ursprungsdaten in Test- und Trainingsdaten (Kreuzvalidierung) oder
(3) mittels Schätzung basierend auf der Residuenquadratsumme ("residual squared sum", RSS)(3), hier im Vergleich zu o.g. SPSE:

$$RSS^{(M)} := \sum_{i=1}^{n} E(Y_i - \hat{Y}_i{}^{(M)})^2$$

$$SPSE^{(M)} := \sum_{i=1}^{n} E(Y_{i+n} - \hat{Y}_i{}^{(M)})^2$$

Es kann gezeigt werden, dass RSS den Wert von SPSE systematisch unterschätzt, dass diese Unterschätzung jedoch behoben werden kann, indem für alle Modelle die Varianzschätzung aus dem maximalen Modell verwendet wird[Sch19]:

$$SPSE^{(M)} := RSS^{(M)} + 2\tilde{\sigma}_{full}^2(k+1)$$

Die Minimierung des SPSE entspricht der Minimierung des Mallow's Cp- Kriteriums, das für die folgenden Analysen getestet werden soll. Dabei gilt:

$$Cp^{(M)} = \frac{1}{\sigma_{full}^2}\sum_{i=1}^{n}(y_i - \hat{y_i}^{(M)})^2 - n + 2(k+1)$$

3.4 Modellselektion

Sei Λ die Menge der erhobenen Wellenlängen in der Designmatrix. Das erste Ziel dieser Arbeit ist es, diejenigen Wellenlängen $\lambda \in \Lambda$ herauszufinden, deren Reflektionswert $\delta(\lambda)$ der Nahinfrarot-Spektroskopie einen Einfluss auf den Stickstoffgehalt des Untergrunds hat. Es wurden mehrere Selektionsverfahren verglichen, die auf zwei Annahmen basieren:
(1) Die erste Ableitung $\delta(\lambda)'$ der Wellenlängen zeigt an, wie groß die Veränderungen der Reflektionswerte sind. Da bei der großen Anzahl an potenziellen Einflussparametern eine Vorauswahl schwierig ist, wurden also diejenigen mit auffälligem mathematischen Verhalten ausgewählt: (a) diejenigen Reflexionen, deren 1. Ableitung über einem Grenzwertes ε_1 lag oder (b) diejenigen, deren Werte unterhalb eines Schwellenwertes lagen.
(2)Darüber hinaus wurden diejenigen Wellenlängen ausgewählt, die einen hohen Informationsgehalt,bzw. Variablilität haben. Zur Berechnung der Variablilität wurden wiederum unterschiedliche Verfahren angewandt. In Modell (a) wurde die Differenz zwischen dem minimalen und dem maximalen Ableitungswert normaiert auf den Mittelwert, während (b) und (c) ohne Normierung arbeiten. In (b) wird lediglich zusätzlich der Betrag verwendet:

(a) $var(\delta(\lambda)) = \frac{|max(\delta(\lambda)') - min(\delta(\lambda)')|}{mean((\delta(\lambda)'))}$
(b) $var(\delta(\lambda)) = |max(\delta(\lambda)') - min(\delta(\lambda)')|$
(c) $var(\delta(\lambda)) = max(\delta(\lambda)') - min(\delta(\lambda)')$

Der Vergleich zwischen den Modellen ergab, dass Mallow's Cp und der zugehörige SPSE am kleinsten sind, wenn wir diejenigen Reflexionen verwenden, deren erste Ableitung und Variablilität über ε_1, bzw. ε_2 liegen. (1b) und (2b), oder:

$$\Delta(\lambda)_{select} = \{(\delta(\lambda)') \in \Lambda | (\delta(\lambda)') > \varepsilon_1 \wedge \\ |max(\delta(\lambda)') - min(\delta(\lambda)')| > \varepsilon_2\}$$

Die so ausgewählten Wellenlängen wurden als Maximalmodell in der Berechnung von Mallow's Cp gegeben. Da immer noch sehr viele Variablen im Modell sind, wurde die performantere Rückwärtsselektion für die konkrete Berechnung verwendet.

3.5 SPSE-Vergleich

Das zweite Ziel dieser Arbeit ist der Vergleich des SPSE-Wertes für das aus 3.4 gewonnene Idealmodell und der SPSE Schätzung über Mallows Cp, welches auf zufällig gezogene Spektren angewandt wurde. Zunächst

4 Implementation

5 Calibration

5.1 Model Selection

SANN returned as minimum values for the respective models

$$C_{\mathrm{p}}^{(\mathrm{SOC})} = -17.46$$
$$C_{\mathrm{p}}^{(\mathrm{N})} = -28.51$$
$$C_{\mathrm{p}}^{(\mathrm{pH})} = -57.76$$

Figures **??**, **??** and **??** in appendix **??** show

5.2 Goodness of Fit

6 Simulation - Schätzgenauigkeit von Mallow's Cp

Die hier vorgestellte Untersuchung soll zeigen, ob und inwiefern sich Mallow‘s Cp als Schätzwert für den SPSE an diesen annähert, wenn sich der Umfang der Stichprobe verändert. Genauer gesagt ziehen wir ein Sample von oben erstelltem Modell, zu dem die exakte Berechnung des SPSE für alle von uns gewählten Stichprobengrößen (z.B. n = [5, 10, 25, 50, 100, 200, 500]) möglichist. Der ermittelte SPSE Wert dient uns im späteren Vergleich als Ground Truth. Für jeden Stichprobenumfang ermitteln wir zudem den Wert von Mallow‘s Cp, der ein Schätzwert des ursprünglichen SPSE darstellt. Ermittelt werden soll dabei, inwiefern sich der Umfang der Stichprobe auf die Abweichung des Schätzers von der Ground Truth (SPSE) auswirkt.

7 Conclusion

Using Mallow's C_{p} criterion, we calibrated three predictive models for the soil parameters $p^{(\mathrm{SOC})}$, $p^{(\mathrm{N})}$ and pH. To construct ...

References

[AHJ10] Lidia Esteve Agelet and Charles R Hurburgh Jr. A tutorial on near infrared spectroscopy and its calibration. *Critical Reviews in Analytical Chemistry*, 40(4):246–260, 2010.

[PDD+13] C Poeplau, A Don, M Dondini, J Leifeld, R Nemo, J Schumacher, N Senapati, and M Wiesmeier. Reproducibility of a soil organic carbon fractionation method to derive rothc carbon pools. *European Journal of Soil Science*, 64(6):735–746, 2013.

[Sch19] Jens Schumacher. Skript zur vorlesung statistische verfahren im wintersemester 2018/19, 2019.

[Umw17] Umweltbundesamt. Stickstoff, 2017.

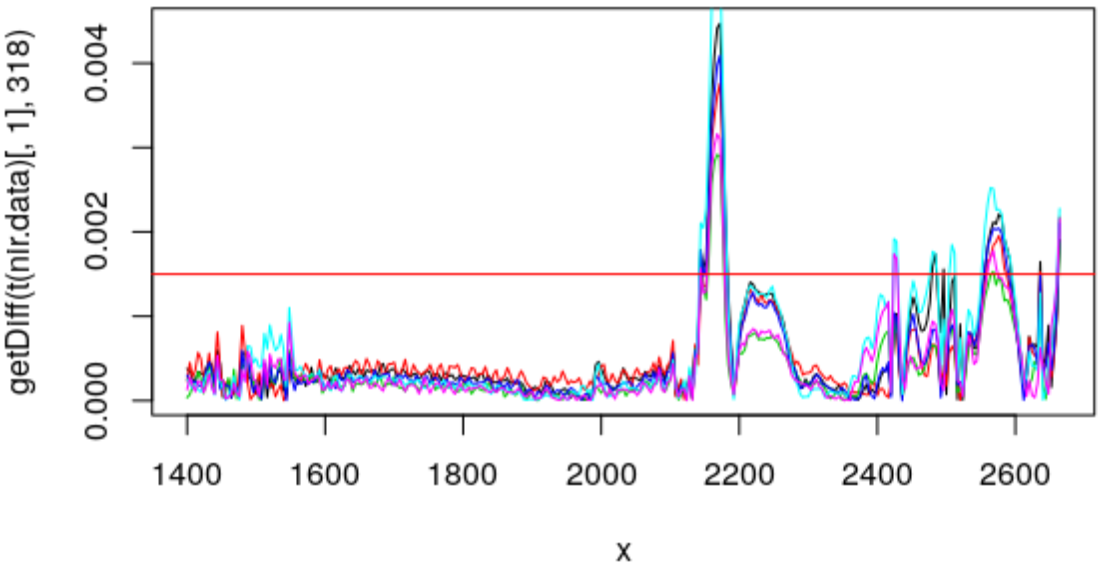
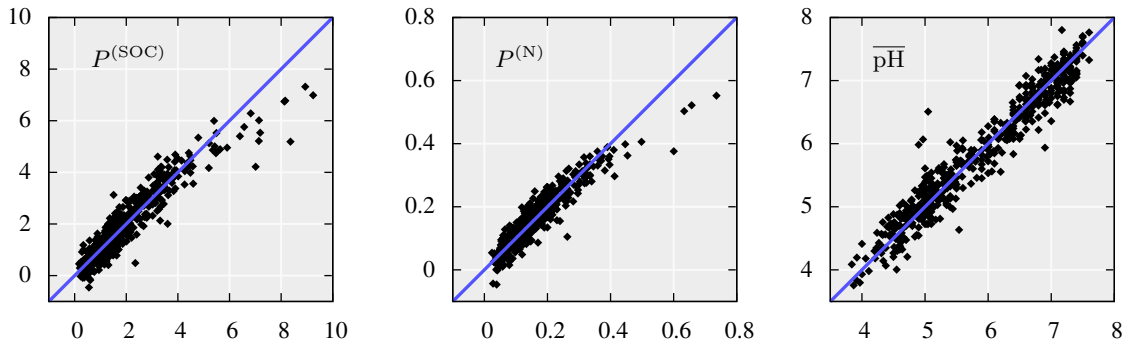


Figure 1: Ableitungen der Wellenlängen von fünf zufälligen Spektren



(a) $\hat{p}^{(\mathrm{SOC})} \sim p^{(\mathrm{SOC})}$ $(R^2)^{(\mathrm{SOC})} = 0.871$

(b) $\hat{p}^{(\mathrm{N})} \sim p^{(\mathrm{N})}$ $(R^2)^{(\mathrm{N})} = 0.861$

(c) $\widehat{\mathrm{pH}} \sim \mathrm{pH}$ $(R^2)^{(\mathrm{pH})} = 0.940$

Figure 2: Correlation diagrams plotting $\hat{y}$ on y and the blue line representing the id

A R Source Code

Statutory Declaration

We herewith declare that ...

Jena, 25th of August 2016

Kazimir Menzel Markus Pawellek