Martina Karoline RESCH, BSc.

Analyse und Modellierung des Chemisch Nickel/Sudgold Prozesses in der Leiterplattenfertigung

MASTERARBEIT

zur Erlangung des akademischen Grades

Diplom-Ingenieur/in

Masterstudium Operations Research und Statistik

eingereicht an der

Technischen Universität Graz

Betreuer/in:
Univ.-Prof. Dipl.-Ing. Dr.techn. Ernst STADLOBER

Institut für Statistik

Graz, September 2015

Diese Arbeit wurde in Kooperation mit der Firma AT&S Austria Technologie & Systemtechnik AG[1] in Leoben, Hinterberg verfasst.

[1] http://www.ats.net/de

EIDESSTATTLICHE ERKLÄRUNG

AFFIDAVIT

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe. Das in TUGRAZonline hochgeladene Textdokument ist mit der vorliegenden Masterarbeit identisch.

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present master's thesis.

Datum/Date	Unterschrift/Signature

Danksagung

Zuallererst möchte ich mich herzlich bei Herrn Professor DI Dr.techn. Ernst Stadlober für die kompetente Betreuung und Unterstützung während der Diplomarbeit und die vielen wertvollen Ratschläge bedanken.

Herzlichst bedanke ich mich auch bei der Firma AT&S, insbesondere bei Technikvorstand Herrn Ing. Heinz Moitzi, der mir diese Arbeit ermöglicht hat. Im Speziellen gilt ein großes Dankeschön DI Thomas Krivec für die freundliche Betreuung und die vielen Stunden voller Engagement und Begeisterung. Ebenso möchte ich mich bei dem gesamten Projektteam bedanken. Danke Doris, für deinen Enthusiasmus und das unermüdliche Zurechtrücken der Ergebnisse. Danke Hans-Peter, für das Vermitteln deines Prozesswissens. Danke Manfred, für dein Mitdenken und Einbringen. Danke Irene, Robert und Evelyn für die zahlreichen Messergebnisse. Außerdem bedanke ich mich bei allen Kollegen in der R&D, die mich freundlich in ihrer Arbeitsgruppe aufgenommen haben.

Mein besonderer Dank gilt meinem Freund Sven und meiner Familie, insbesondere meinen Eltern, die mir das Studieren ermöglicht haben, und meiner Schwester Susi – ohne euch und eure Hilfe in allen Bereichen hätte ich das nicht geschafft.

Kurzfassung

Trotz der immer größer werdenden Menge an vorhandenen Daten werden Entscheidungen oft ohne vorherige Datenanalyse und daraus gewonnenem Wissen getroffen. Die Statistik bietet die Möglichkeit, gesammelte Daten zu untersuchen und damit die Informationen in den Daten zu nutzen. Welches Potenzial statistische Auswertungen im Hinblick auf Prozessverständnis, Prozessmodellierung und Prozessoptimierung haben können, wird anhand des Chemisch Nickel/Sudgold Prozesses in der Leiterplattenfertigung veranschaulicht.

Diese Arbeit beschäftigt sich mit der Untersuchung der Einflüsse der Prozessparameter auf die mittlere Goldschichtstärke im Chemisch Nickel/Sudgold Prozess. Mit Hilfe von explorativen Analysen werden Strukturen und Abhängigkeiten der Prozessparameter und Schichtstärken untersucht. Daraus wird abgeleitet, welche Prozessparameter eine adäquate Beschreibung der Zielgröße Goldschichtstärke ermöglichen. Mittels multivariaten multiplen Regressionsmodellen werden signifikante Zusammenhänge beschrieben.

Dazu werden unterschiedliche Datensätze verwendet - ein Datensatz für die Vorstudie, ein weiterer für die Hauptstudie und ein dritter zur Validierung der erarbeiteten Regressionsmodelle. Dabei stellt sich heraus, dass sich die beobachteten Prozessparameter zwar zur Beschreibung einzelner Datensätze eignen, aber nur eingeschränkt für eine Vorhersage bei neuen Daten verwendet werden können.

Sämtliche statistische Analysen wurden mit dem Statistik-Programm **R** (Version 3.1.2) durchgeführt.

Diese Arbeit wurde in Kooperation mit der Firma AT&S Austria Technologie & Systemtechnik AG in Leoben, Hinterberg verfasst.

Abstract

Although there are more and more data available these days, some decisions are made without any analysis of them, and therefore without using the gained know-how. Statistics offer the ability to examine collected data and to benefit from the information in the data. The purpose of this thesis is to illustrate the use of statistical methods for improving the process understanding, process modeling and process optimization of the electroless nickel/immersion gold process in printed circuit board production.

This thesis deals with the investigation of the influence of the process parameters on the average gold layer thickness in the electroless nickel/immersion gold process. Exploratory data analysis are used to explore structures and correlations of process parameters and layer thicknesses. From this analysis it is deduced which process parameters are useable for an adequate description of the response gold layer thickness. Significant correlations are described by means of multivariate multiple regression models.

Therefore, three different data sets are analysed. On the one hand, a data set for the preliminary study is used, on the other hand, another data set for the main study is analysed. A third one is used for the validation of the developed regression models. The results show that the observed process parameters are suitable for the description of single data sets, but they are of limited use for prediction.

All statistical analyses were performed using the statistics program **R** (Version 3.1.2).

This master thesis is written in collaboration with AT&S Austria Technologie & Systemtechnik AG in Leoben, Hinterberg.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AIC	Akaike´s Informationskriterium
AIC$_c$	Akaike´s Informationskriterium (korrigiert)
BIC	Bayes´sches Informationskriterium
CSV	Comma-Separated Values
EDA	Explorative Datenanalyse
ENIG	Electroless Nickel Immersion Gold
IQR	Interquartiler Bereich
MTO	Metal Turn Over
Q-Q-Plot	Quantil-Quantil-Plot
SSE	Error Sum of Squares (Fehler-Quadratsumme)
SSR	Regression Sum of Squares (Regressions-Quadratsumme)
SST	Total Sum of Squares (Totale Quadratsumme)

1 Einleitung

Heutzutage werden immer mehr Daten erfasst und gespeichert. Diese stellen einen wesentlichen Ausgangspunkt für einen Entscheidungsprozess dar. Trotzdem werden Entscheidungen getroffen, ohne zuvor vorhandenes Datenmaterial zu analysieren und daraus gewonnenes Wissen in den Entscheidungsprozess einfließen zu lassen. Die Statistik bietet die Möglichkeit, gesammelte Daten zu untersuchen und damit die Informationen in den Daten zu nutzen, um Entschlüsse nicht willkürlich sondern wissensbasiert zu fassen. Dies wird am Beispiel des Chemisch Nickel/Sudgold Prozesses (Electroless Nickel Immersion Gold (ENIG) - Prozess) in der Leiterplattenfertigung veranschaulicht.

Ziel dieser Arbeit ist die Untersuchung der Einflüsse der Prozessparameter auf die Goldschichtstärke im ENIG-Prozess mit Hilfe von explorativen Analysen und die Modellierung des Prozesses mittels multivariaten multiplen Regressionsmodellen. Damit soll bisher nicht genutztes Datenmaterial statistisch analysiert und modelliert werden, um vorhandenes Prozesswissen zu erweitern und zukünftige Prozessentscheidungen zu erleichtern. Darüber hinaus soll die Prozessführung optimiert werden, indem der Edelmetallverbrauch minimiert wird.

Zwei zentrale Fragen stellen sich:

- Welche Prozessparameter haben einen Einfluss auf die Goldschichtstärke?
- Wie sollen die relevanten Prozessparameter eingestellt werden um den Edelmetallverbrauch zu minimieren?

Die Arbeit gliedert sich in folgende Bereiche:

Das Kapitel 2, Die Firma AT&S und der ENIG-Prozess, widmet sich der Firmenbeschreibung und der Einführung in den zu analysierenden ENIG-Prozess.

Das Kapitel 3, Theoretische Grundlagen, beinhaltet eine Zusammenfassung der in dieser Arbeit verwendeten statistischen Bausteine. Die explorative Datenanalyse und die lineare Regression inklusive Regressionsbäumen sind darin kurz beschrieben.

Das Hauptkapitel 4, Praktische Problemlösung, umfasst die praktische Anwendung aller im vorigen Kapitel erläuterten Themen auf den ENIG-Prozess. Als erster Schritt wird eine Vorstudie anhand eines vorhandenen Datensatzes mittels explorativen Analysen durchgeführt. Basierend auf diesen Erkenntnissen wird ein Regressionsmodell für die mittlere Goldschichtstärke erstellt.
Die Hauptstudie basiert auf einem zweiten, optimierten Datensatz. Dieser wird statistisch untersucht und ein Regressionsmodell für die mittlere Goldschichtstärke ausgearbeitet.

Im Folgenden wird das vorgeführte Regressionsmodell anhand eines Validierungsdatensatzes auf seine Effizienz hin überprüft.

Abschließend sind alle wichtigen Ergebnisse in Kapitel 5, Zusammenfassung, dargestellt.

Appendix, Statistische Kenngrößen, beinhaltet die verwendeten mathematischen Grundbegriffe und Definitionen.

2 Die Firma AT&S und der ENIG-Prozess

2.1 Die Firma AT&S

Die Firma AT&S, Austria Technologie und Systemtechnik Aktiengesellschaft, ist Europas größter und weltweit einer der leistungsstärksten Produzenten von hochwertigen Leiterplatten, siehe AT&S AG, 2015b bzw. AT&S AG, 2015c.

Das Unternehmen wurde 1987 in der Steiermark gegründet und ist heute mit rund 8000 Mitarbeiterinnen und Mitarbeitern weltweit tätig. Der Hauptsitz befindet sich in Leoben und die Produktionsstandorte konzentrieren sich in Europa auf Leoben und Fehring in Österreich und in Asien auf Ansan in Korea, Nanjangud in Indien und Shanghai in China. Der Produktionsstandort China wird derzeit mit dem Aufbau eines Werks in Chongqing erweitert. Außerdem zählen zur AT&S neben einem eigenen Vertriebs-, Service- und Designzentrum in Nörvenich in Deutschland zahlreiche internationale Vertriebsbüros, unter anderem in Japan und den USA.

Alle Standorte der AT&S sind auf unterschiedliche Aufgabenbereiche spezialisiert. Die Werke in Österreich zeichnen sich durch kurze Durchlaufzeiten und Spezialanwendungen aus. Die wichtigsten Absatzmärkte dafür befinden sich in Europa und Amerika. Im Gegensatz dazu liegt das Hauptaugenmerk der Fertigung in China auf Großserien für Kunden aus der Mobilkommunikationsindustrie.

Das Kerngeschäft der AT&S ist die Produktion von Leiterplatten. Diese Leiterplatten bilden das „Nervenzentrum“ sämtlicher elektronischer Geräte, vgl. AT&S AG, 2015a. Sie dienen als Trägermaterial elektrischer Verbindungen von Bauelementen und sind Medium für elektrische Signale.

Der Herstellungsprozess von Leiterplatten beinhaltet drei Schwerpunkte:

- den Datenbearbeitungsprozess,
- den Herstellungsprozess und
- den Nachbearbeitungsprozess.

Im Datenbearbeitungsprozess werden Angebote für Kunden anhand von Layout-Daten, Skizzen und Produktanforderungen erarbeitet und für die Fertigung aufbereitet.

Im Herstellungsprozess werden Basismaterial und Kupferfolien mit den strukturierenden Cores verpresst und daraufhin belichtet und geätzt. So entsteht ein Leiterbild. Falls die Leiterplatte aus mehreren Lagen besteht, werden mehrere der entstandenen Lagen verpresst und zum vertikalen Verbinden der Lagen gebohrt. Um die Bohrungen

leitend zu machen, werden sie verkupfert. Die Leiterbahnen der Außenlagen entstehen wiederum durch Belichten und Ätzen.

Im Nachbearbeitungsprozess werden die äußeren Lagen der bereits funktionierenden Leiterplatten mit zwei Schutzschichten überzogen, eine Schutzschicht für die Außenlagen und eine Schutzschicht für die Kontaktstellen. Abschließend werden die Leiterplatten in zwei Testverfahren kontrolliert.

Der in dieser Arbeit analysierte Prozessschritt ist der Chemisch Nickel/Sudgold Prozess innerhalb des Nachbearbeitungsprozesses. Darin findet die Veredelung der Kontaktflächen der Leiterplatten mit einer funktionalen Schutzschicht statt.

2.2 Der ENIG-Prozess

Ein zentraler Prozess in der Leiterplattenherstellung ist der Chemisch Nickel/Sudgold Prozess, kurz ENIG-Prozess, siehe AT&S AG, 2015a. Dieser ist einer der letzten Schritte in der Leiterplattenproduktion.

Im vorangegangenen Lötstopplack-Prozess wird die gesamte Oberfläche einer Leiterplatte unter Aussparung der Kontaktflächen mit einem Lötstopplack überzogen, vgl. Abbildung 2.1 (Schritt 2). Dieser dient als Schutzschicht für die Außenlagen. Anschließend folgt der ENIG-Prozess. Dabei wird eine funktionelle Schutzschicht auf die Kontaktstellen einer bereits funktionstüchtigen Leiterplatte aufgetragen, vgl. Abbildung 2.1 (Schritt 3). Um die ausgesparten aus Kupfer bestehenden Kontaktflächen vor Oxidation zu schützen und die geforderte Haftung für eine spätere Bestückung zu bieten, werden die lackfreien Stellen mit einer Goldschicht veredelt. Gold ist ein Edelmetall und bildet keine Oxidschicht. Daher ist es für die Oberflächenbeschichtung sehr gut geeignet.

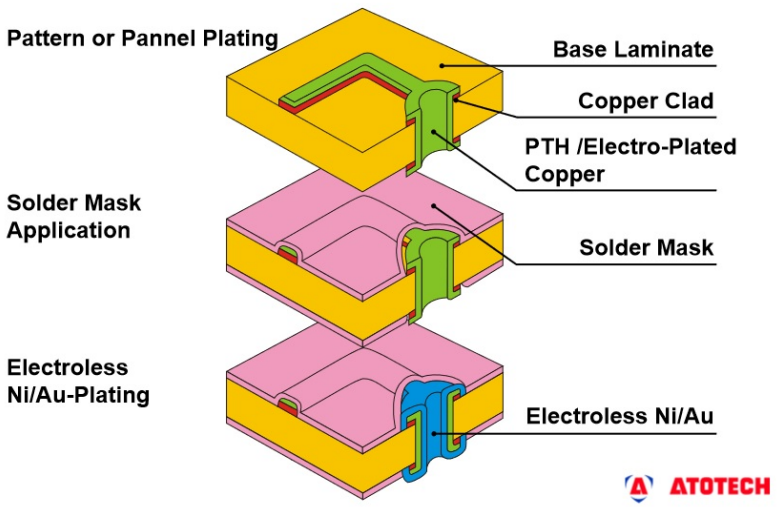

Abbildung 2.1: ENIG-Prozess (Atotech GmbH, 2014).

Die Bezeichnung „chemisch“ erklärt sich dadurch, dass das Abscheiden der funktionellen Oberflächenbeschichtung auf einer rein chemischen Reaktion beruht. In der AT&S basieren 50 Prozent der Oberflächenveredelung auf „Chemisch Nickel/Gold“ und 30 Prozent der Oberflächenveredelung auf „Chemisch Zinn“. Die restlichen 20 Prozent bilden die Verfahren HAL (Hot Air Levelling) und HAL bleifrei sowie OSP (Organic Surface Protection).

Der ENIG-Prozess ist ein vertikales Tauchverfahren, deren Anlage in Aktivstationen und Spülstationen aufgeteilt werden kann. Der Prozess dauert in etwa 2.5 Stunden. Die Aktivstationen sind in Abbildung 2.2 dargestellt. Zwischen je zwei aufeinanderfolgenden Aktivstationen befinden sich Spülstationen.

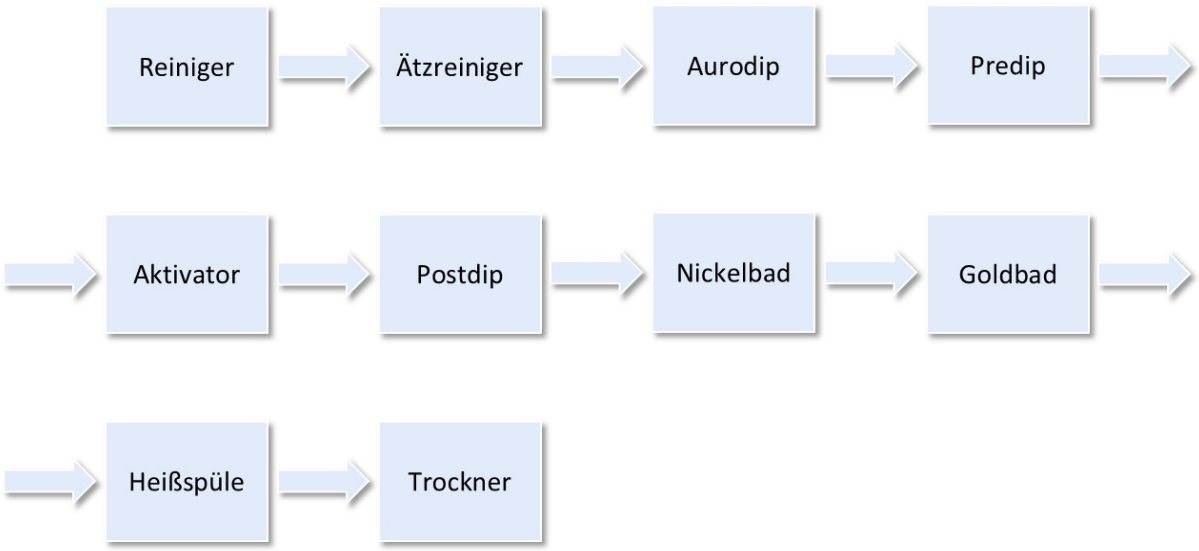

Abbildung 2.2: Aktivstationen des ENIG-Prozesses.

Die Aktivstationen lassen sich folgendermaßen beschreiben (Atotech GmbH, 2014):

- **Reiniger**
  Im Reiniger werden Kupferoxide und Fette (Fingerabdrücke, Ölflecken,...) entfernt. Dies ist nötig um ein vollständiges Benetzen zu gewährleisten.
- **Ätzreiniger**
  Im Ätzreiniger werden ca. 1.5 μm bis 2.5 μm der Kupferschicht abgeätzt. Die dabei entstehende Mikro-Rauigkeit wird für den Folgeprozess benötigt. Zudem werden Lötstopplack-Rückstände entfernt.
- **Aurodip**
  Der Aurodip besteht aus einer heißen Säurespüle.
- **Predip**
  Im Predip findet eine Desoxidation und Spülung der Kupferschicht statt. Dadurch wird die Oberfläche für den Aktivator konditioniert.
- **Aktivator**
  Im Aktivator wird ein „Monolayer“ Palladium geplatet. Diese Palladiumschicht initiiert die Nickel-Abscheidung.

- **Postdip**
  Der Postdip ist eine saure Wasserspüle und dient als „Verschleppungsschutz“ für Palladium-Chemien.
- **Nickelbad**
  Im Nickelbad wird eine Nickelschicht mit einer Schichtstärke zwischen 3 μm und 6 μm abgeschieden.
- **Goldbad**
  Im Goldbad wird eine dünne Goldschicht mit einer Schichtstärke zwischen 0.06 μm und 0.12 μm abgeschieden, welche die Nickelschicht vor Oxidation schützt.
- **Heißspüle**
  In der Heißspüle werden restliche Prozess-Rückstände entfernt.
- **Trockner**
  Im Trockner wird die Leiterplatte getrocknet.

Durch diesen Prozess entsteht eine Leiterplatte mit einer universell einsetzbaren Oberfläche, die sich durch eine lange Haltbarkeit auszeichnet.

3 Theoretische Grundlagen

Dieses Kapitel gibt einen kurzen Einblick in die theoretischen statistischen Grundlagen, die in der vorliegenden Arbeit verwendet werden. Vorwiegend folgt dieses Kapitel den Ausführungen in Burkschat, Cramer und Kamps, 2012, Fahrmeir, Kneib und Lang, 2009, Friedl, 2008, und Stadlober, 2008. Für weitere Informationen wird auf die Literaturliste verwiesen.

3.1 Explorative Datenanalyse

Die Explorative Datenanalyse (EDA) stellt eine Art Voruntersuchung der Daten dar. Mittels spezifischer Verfahren können Muster, Auffälligkeiten und Unregelmäßigkeiten in ein- und mehrdimensionalen Datensätzen entdeckt werden, um einen Überblick davon zu bekommen, wie sich Daten in Bezug auf Streuung, Symmetrie, Konzentration, Ausreißer etc. verhalten. Dadurch können Ideen gewonnen werden, die mittels Hypothesentests auf ihre Gültigkeit hin überprüft werden können, siehe Akkerboom, 2012.

Univariate Analysen einzelner Variablen und multivariate Analysen mehrerer Variablen stellen die Grundlage dar. So wird zuerst jede Variable für sich selbst untersucht um anschließend Zusammenhänge zwischen den Variablen herauszufinden, vgl. Cleff, 2011.

Die Untersuchung einer Variablen kann einerseits mit statistischen Kennzahlen von Lage und Streuung geschehen, als auch mit anschaulichen Grafiken. Allein die Kenntnis von einigen Werten einer Verteilung gibt meist keine vollständige Idee über die gesamte Verteilung. Anhand verschiedenster Grafiken, wie zum Beispiel Stängel-Blattdiagrammen, Boxplots, Histogrammen oder Quantil-Quantil-Plots ist die Verteilung leichter erfassbar und besondere Strukturen können entdeckt werden. Für multivariate Auswertungen werden unter anderem Scatterplots als grafisches Analyseelement herangezogen, siehe Kohn und Öztürk, 2013.

3.1.1 Stängel-Blattdiagramm

Ein Stängel-Blattdiagramm ist eine gegliederte Liste für aufsteigend sortierte Werte. Alle Beobachtungen werden dafür in zwei Teile aufgespalten. Der eine Teil besteht aus der letzten inhaltlich relevanten Ziffer des beobachteten Wertes und bildet die Blattziffer. Der andere Teil wird aus den verbleibenden Ziffern gebildet und als Stammzahl bezeichnet, siehe Wermuth und Streit, 2007.

Ein Beispiel dafür liefert die linke Grafik in Abbildung 3.1. Der Stamm ist stehend dargestellt und wird aus den Zahlen 64, 66,..., 98 gebildet. Die Blattziffern ergeben sich aus den zugehörigen Kommastellen und sind der entsprechenden Stammzahl zugeordnet.

Anhand dieses Diagramms sind Extremwerte, der Streubereich der beobachteten Werte, die Stammzahl mit den meisten Blättern und häufig vorkommende Blätter leicht zu erkennen. Da alle Daten abgebildet werden, entsteht kein Datenverlust, Stadlober, 2006.

3.1.2 Boxplot

Die beobachtete Stichprobe $(x_1, \ldots, x_n)$ eines Merkmals X sei gegeben. Der Boxplot (siehe rechte Grafik in Abbildung 3.1) stellt Median, 1. und 3. Quartil, sowie Minimum und Maximum grafisch dar und ist damit ein Werkzeug um die Lage und Streuung von Variablen zu beurteilen.

```
The decimal point is at the |

  64 | 588
  66 | 89
  68 | 0416777
  70 | 00244467780345568
  72 | 00112256689011115556669999
  74 | 33344455667789334467889 99
  76 | 12233455568889901223344679
  78 | 01133445556788911112222334444455667789
  80 | 003335677700037789 9
  82 | 0013344556680345568888
  84 | 56678901113569
  86 | 116814
  88 | 1226
  90 | 247711278
  92 | 0261
  94 | 69
  96 | 38
  98 | 03888
```

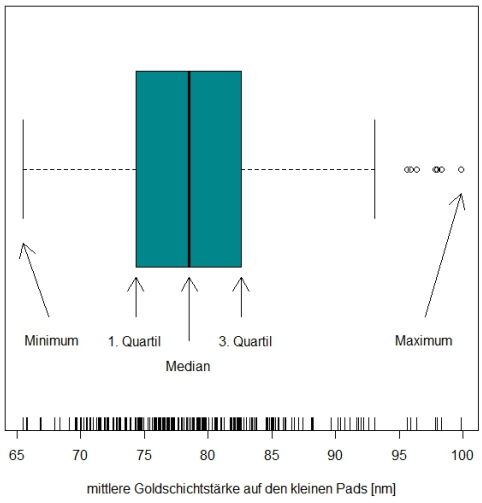

Abbildung 3.1: Stängel-Blattdiagramm und Boxplot der mittleren Goldschichtstärke auf den kleinen Pads (ENIG.neu).

Die Grafik besteht aus einer Box, welche die mittleren 50% der Daten (definitionsgemäß zwischen dem 1. und 3. Quartil) darstellt, und zwei Tails, welche die Werte außerhalb symbolisieren. Die horizontale Linie innerhalb der Box gibt die Lage des Median wieder. Liegt der Median in der Mitte der Box, sind die zentralen Daten symmetrisch. Die Länge der Tails ist definiert als (Kohn und Öztürk, 2013)

$$\begin{aligned} Tail_{min} &= \min_{1 \le i \le n} \{x_i | x_i \ge q_{0.25} - 1,5\,(q_{0.75} - q_{0.25})\} \text{ und} \\ Tail_{max} &= \max_{1 \le i \le n} \{x_i | x_i \le q_{0.75} + 1,5\,(q_{0.75} - q_{0.25})\}. \end{aligned} \tag{3.1}$$

Dabei bezeichnen $q_{0.25}$ bzw. $q_{0.75}$ das 1. bzw. 3. Quartil. Werte außerhalb des Box- und Tail-Bereichs werden Ausreißer genannt. Sie sind im Vergleich zur „breiten Masse“ verhältnismäßig klein oder groß.

3.1.3 Beeswarm-Boxplot

Der Beeswarm-Boxplot, Abbildung 3.2, ist ein Spezialfall des Boxplots. Mit der **R**-Funktion *beeswarm* des Pakets *beeswarm* können in einen Boxplot zusätzlich einzelne Punkte kategorisch gefärbt werden, siehe Eklund, 2015. So können Zusammenhänge und Muster zwischen zwei Variablen erkannt werden. Für weitere Informationen wird auf die Ausführungen in Eklund, 2015, verwiesen.

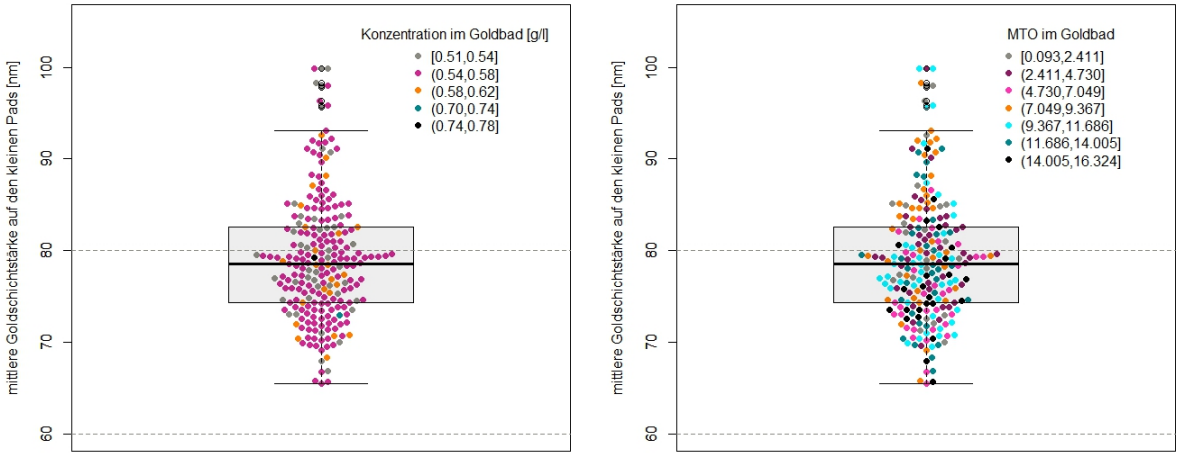

Abbildung 3.2: Beeswarm-Boxplots (ENIG.neu). Der Boxplot zeigt die Verteilung der mittleren Goldschichtstärke der kleinen Pads. Die Bienenschwärme stellen den Zusammenhang mit der Konzentration bzw. des Metal Turn Overs (MTO) im Goldbad dar.

3.1.4 Histogramm

Eine andere Art die Verteilung der Werte darzustellen, ist durch das Histogramm gegeben, siehe linke Grafik in Abbildung 3.3. Dabei werden die vorliegenden n Werte $x_1, \ldots, x_n$ der Stichprobe in m äquidistante Klassen zusammengefasst. Die Häufigkeiten der einzelnen Klassen werden im Histogramm abgebildet.

Die geordneten Daten $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ bilden die Grundlage. Der gesamte Bereich der Stichprobe $\left[x_{(1)}, x_{(n)}\right]$ wird in k Klassen, die alle die gleiche Breite b haben, aufgeteilt (Stadlober, 2006)

$$\left[c_0 \leq x_{(1)}, c_1\right), \left[c_1, c_2\right), \ldots, \left[c_{k-1}, c_k \geq x_{(n)}\right), \text{ mit } c_j = c_0 + jb,\ j = 0, \ldots, k.$$

Das Histogramm setzt sich aus k Rechtecken mit Breite b und Länge h_j zusammen. Diese werden über den zugehörigen Klassen $\left[c_{j-1}, c_j\right)$ abgebildet

$$h(x) = \begin{cases} h_j, & \text{falls } c_{j-1} \leq x < c_j,\ j = 1, \ldots, k;\ c_j - c_{j-1} = b, \\ 0, & \text{sonst.} \end{cases} \tag{3.2}$$

Dabei wird h_j, je nach gewünschter Skalierung, als absolute Häufigkeit, relative Häufigkeit oder Dichteschätzer definiert.

Bemerkung

- Unterschiedliche Klassenbreiten wirken sich auf das Erscheinungsbild des Histogramms aus, vgl. Toutenburg u. a., 2009.
- Der Bezug zu den Originaldaten geht durch die Klassifizierung verloren, vgl. Stadlober, 2008.

3.1.5 Quantil-Quantil-Plot (Q-Q-Plot)

Die Anpassung einer theoretischen Verteilung F an die beobachtete Stichprobe kann unter Zuhilfenahme von Q-Q-Plots überprüft werden, siehe rechte Grafik in Abbildung 3.3. Dabei werden die geordneten Werte der Stichprobe x_i gegen die theoretischen Quantile $F^{-1}\left(\frac{i}{n+1}\right)$, $i = 1, \ldots, n$ aufgetragen. Folgend kann die Abweichung von der Referenzgeraden, einer Gerade mit einer 45° Steigung, beurteilt werden.

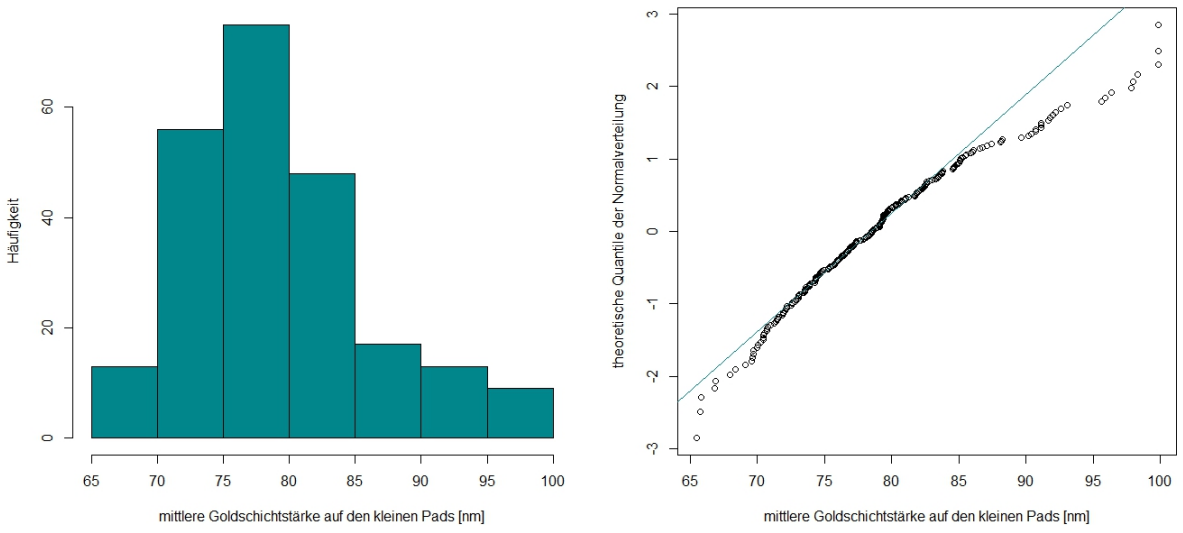

Abbildung 3.3: Histogramm und Q-Q-Plot der mittleren Goldschichtstärke auf den kleinen Pads (ENIG.neu).

3.1.6 Scatterplot

Die beobachtete paarige Stichprobe (x_i, y_i), $i = 1, \ldots, n$, zweier Merkmale X und Y sei gegeben. Abbildung 3.4 gibt die Punkte der zwei Merkmale in einem Koordinatensystem wieder.

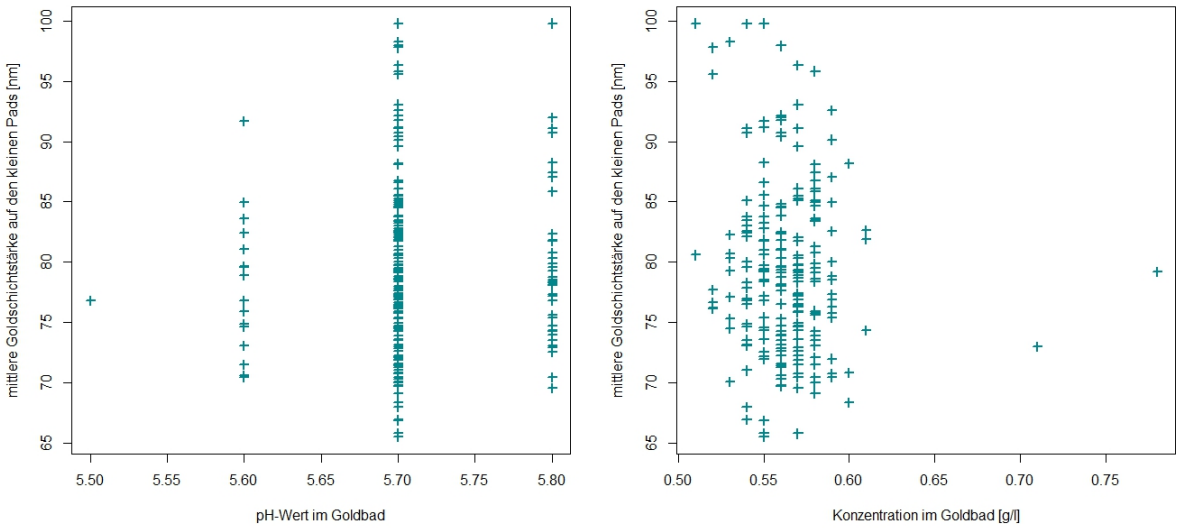

Abbildung 3.4: Scatterplot der mittleren Goldschichtstärke auf den kleinen Pads gegen den pH-Wert im Goldbad bzw. die Konzentration im Goldbad (ENIG.neu).

Auf der x-Achse sind die Ausprägungen des einen Merkmals aufgetragen, während auf der y-Achse die Ausprägungen des zweiten Merkmals aufgetragen sind. Das Ergebnis ist eine Punktwolke, die als Scatterplot (Streudiagramm) bezeichnet wird, vgl. Pruscha, 2006.

Bemerkung

- Ob ein Zusammenhang zwischen den beiden Merkmalen besteht, kann meistens schon anhand dieser Grafik beurteilt werden, siehe Stadlober, 2006.
- Die beobachteten Werte der Zielvariablen werden auf der y-Achse aufgetragen und die Werte des Prädiktors auf der x-Achse. Für gleichgestellte Variablen kann die Zuordnung der Achsen frei gewählt werden, siehe Wermuth und Streit, 2007.

3.1.7 Korrelation

Falls im Scatterplot festgestellt wird, dass für größere Werte der Stichprobe X auch Y größere Werte aufweist oder dass für wachsende Werte von X kleinere Werte von Y auftreten, scheint es plausibel eine Abhängigkeit zwischen den Werten zu vermuten. Die Stärke des linearen Zusammenhangs gibt der Korrelationskoeffizient an.

Die beobachteten n paarigen Daten der Variablen X und Y werden mit $(x_1, y_1), \ldots, (x_n, y_n)$ bezeichnet. Der Korrelationskoeffizient r_{xy} der Stichprobe von (X, Y) ist definiert als

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \tag{3.3}$$

X und Y sind

- positiv korreliert, falls $r_{xy} > 0$,
- unkorreliert, falls $r_{xy} = 0$, und
- negativ korreliert, falls $r_{xy} < 0$.

Bemerkung

- Burkschat, Cramer und Kamps, 2012, teilt die Stärke des Korrelationskoeffizienten
  - in eine schwache Korrelation $0 \leq |r_{xy}| < 0,5$ und
  - in eine starke Korrelation $0,8 \leq |r_{xy}| \leq 1$ auf.
- Mit dem Korrelationskoeffizienten können keine Kausalzusammenhänge identifiziert werden. Dies kann nur mit theoretischen Überlegungen geschehen, vgl. Kohn, 2005, und Kronthaler, 2014.
- Nichtlineare Zusammenhänge können mittels Korrelationskoeffizienten nicht erkannt werden. Deswegen ist die Betrachtung der zugehörigen Scatterplots wesentlich, siehe Kohn und Öztürk, 2013.

3.2 Lineare Regression

„A first (...) principle is that models are wrong; some, though, are more useful than others and we should seek those. A second principle (...) is not to fall in love with one model to exclusion of alternatives. A third principle recommends thorough checks on the fit of a model to the data, for example by using residuals and other statistics derived from the fit for outlying observations and so on."
(McCullagh und Nelder, 1989)

Die Regressionsanalyse dient der Modellierung des Zusammenhangs einer Variable Y, auch Zielvariable, Response oder abhängige Variable genannt, und einer oder mehrerer Prädiktoren bzw. unabhängigen Variablen $X_1, \ldots, X_{p-1}$ (Faraway, 2004). Dabei wird zwischen verschiedenen Arten der Regressionsanalyse unterschieden

- $p = 2$: einfache lineare Regression,
- $p > 2$: multiple lineare Regression und
- $p > 2$ und mehrere Zielvariablen $Y_1, \ldots, Y_m$, $m \geq 2$: multivariate multiple lineare Regression.

Unabhängig von der Art der Regressionsanalyse, können unterschiedliche Ziele verfolgt werden

- die Beschreibung der Datenstruktur,
- das Verständnis der Beziehung zwischen der Zielvariablen und den Prädiktoren oder
- die Vorhersage zukünftiger Beobachtungen.

In dieser Arbeit wird vorwiegend auf das Verständnis der Beziehung zwischen der Zielvariablen und den Prädiktoren und die Vorhersage zukünftiger Beobachtungen Wert gelegt.

3.3 Multiple lineare Regression

Mit Hilfe der linearen Regressionsanalyse wird ein linearer Zusammenhang zwischen einer Zielvariablen Y und einer oder mehreren Prädiktoren $X_1, \ldots, X_{p-1}$ modelliert, vgl. Kohn, 2005. Dabei wird angenommen, dass für alle einzelnen Werte der Prädiktoren $X_1, \ldots, X_{p-1}$ die Zielvariable Y zufällig ist und deren Erwartungswert von $X_1, \ldots, X_{p-1}$ abhängt. Die lineare Regression, für die mehr als ein Prädiktor zur Erklärung herangezogen wird, wird als multiple lineare Regression bezeichnet. Die Regressionsfunktion ist dabei linear in den Parametern.

Allgemein ist $(x_{i1}, x_{i2}, \cdots, x_{i,p-1}, y_i),\ i = 1, 2, \ldots, n$ die Menge der gegebenen Daten. Das Regressionsmodell ist definiert als

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i, \qquad i = 1, \ldots, n, \tag{3.4}$$

bzw.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.5}$$

mit Zielvariablenvektor $\mathbf{y} = (Y_1, \ldots, Y_n)^T$, Parametervektor $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})^T$ und zufälligem Fehlervektor $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, sowie der $n \times p$ Designmatrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix}.$$

Bemerkung

- Die erste Spalte der Designmatrix muss aus lauter Einsern bestehen, um den Intercept im Modell zu berücksichtigen (Fahrmeir, Kneib und Lang, 2009).

Dabei werden einige Modellannahmen getroffen

- $\epsilon_i \overset{iid}{\sim} \text{Normal}(0, \sigma^2),\ i = 1, \ldots, n \Leftrightarrow \boldsymbol{\epsilon} \sim \text{Normal}(\mathbf{0}, \sigma^2\mathbf{I}) \Leftrightarrow \mathbf{y} \sim \text{Normal}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$,
- $\beta_0, \beta_1, \ldots, \beta_{p-1}$ und σ^2 sind unbekannte Parameter und
- $x_{ij},\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p-1$, sind bekannte Konstanten.

Folgende Konsequenzen lassen sich, laut Friedl, 2008, daraus ableiten

- Die Zielvariable Y_i lässt sich als Summe eines konstanten Terms $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$ und eines zufälligen Terms ϵ_i darstellen. Daher ist die Zielvariable Y_i eine Zufallsvariable.
- Die zufälligen Fehler $\epsilon_i,\ i = 1, \ldots, n$, sind unabhängig $\Rightarrow$ die Zielvariablen Y_i, $i = 1, \ldots, n$, sind unabhängig.
- $\mathrm{E}(Y_i) = \mathrm{E}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} \Leftrightarrow \mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$
- $\mathrm{Var}(Y_i) = \mathrm{Var}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i) = \mathrm{Var}(\epsilon_i) = \sigma^2$. Das bedeutet, alle y_i haben die gleiche, konstante Varianz.

Bemerkung

- Die einfache lineare Regression ist ein Spezialfall der multiplen linearen Regression, indem nur ein Prädiktor verwendet wird.
- In einem linearen Modell sind die Koeffizienten $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{p-1})$ linear. Die Prädiktoren müssen nicht linear sein. Die Verwendung von linearen Modellen erscheint unter Umständen als sehr einschränkend, aber mittels Transformationen und Kombinationen der Prädiktoren sind lineare Modelle sehr flexibel, siehe Faraway, 2004.
- Bei der multiplen linearen Regression kann es vorkommen, dass Prädiktoren untereinander korreliert sind. Da hoch korrelierte Variablen dasselbe erklären, ist die Verwendung nur einer dieser Variablen notwendig.
  Außerdem stellt die Verwendung hoch korrelierter Prädiktoren in der Regressionsanalyse ein Problem dar, welches als Multikollinearität bezeichnet wird. Dieses führt zu einer schlechteren Schätzung der Parameter β_i.

3.3.1 Parameterschätzungen

Für das Regressionsmodell sind $p + 1$ Parameter, $\beta_0, \ldots, \beta_{p-1}$ und σ^2, unbekannt. Anhand der gegebenen Beobachtungen müssen diese geschätzt werden, vgl. Kohn, 2005.

Schätzung des Parametervektors β

Die Schätzung von $\boldsymbol{\beta}$ soll so sein, dass der Fehler zwischen $\mathbf{X}\boldsymbol{\beta}$ und $\mathbf{y}$ möglichst gering ist. In Abbildung 3.5 ist die beste Wahl von $\boldsymbol{\beta}$ geometrisch dargestellt. Dabei wird der Vektor $\mathbf{y}$ orthogonal auf den von $\mathbf{X}$ aufgespannten Modellraum projiziert. Die Residuen $\mathbf{e} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ ergeben sich als Differenz zwischen den beobachteten Werten und den geschätzten Werten.

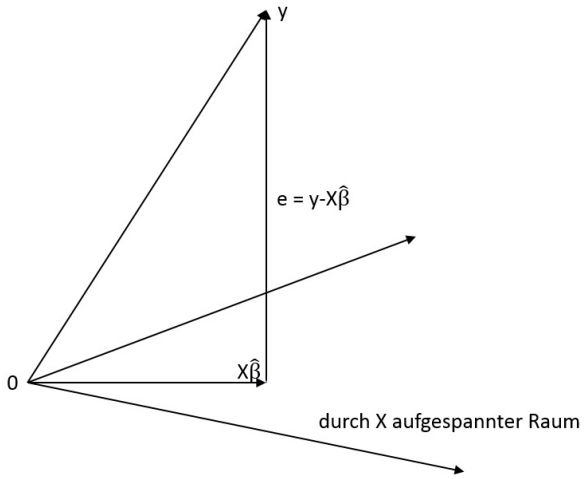

Abbildung 3.5: Geometrische Darstellung der Schätzung von $\boldsymbol{\beta}$. Der Vektor $\mathbf{y}$ wird orthogonal auf den von $\mathbf{X}$ aufgespannten Modellraum projiziert.

Die Schätzung der Regressionskoeffizienten erfolgt mittels Kleinster Quadrate Methode. Der beste Schätzer von $\boldsymbol{\beta}$ ist per Definition jener Schätzer, der die Summe der quadrierten Fehler minimiert. Anders ausgedrückt, der kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}$ minimiert die Gleichung

$$\text{SSE}(\boldsymbol{\beta}) = \sum \epsilon_i^2 = \boldsymbol{\epsilon}^T \boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}). \tag{3.6}$$

Durch Minimierung der Gleichung (3.6) ergibt sich für den Schätzer $\hat{\boldsymbol{\beta}}$ folgender Ausdruck

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T\mathbf{y} \sim \text{Normal}\left(\boldsymbol{\beta}, \sigma^2 \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right). \tag{3.7}$$

Unter Verwendung der Hat-Matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ kann folgende Beziehung dargestellt werden

$$\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}. \tag{3.8}$$

Die Hat-Matrix ist demnach die orthogonale Projektion von $\mathbf{y}$ auf den von $\mathbf{X}$ aufgespannten Modellraum.

Eine andere Möglichkeit $\boldsymbol{\beta}$ zu schätzen ist die Maximum-Likelihood Schätzung. Dabei wird die Loglikelihood Funktion maximiert

$$\log \text{f}(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\text{SSE}(\boldsymbol{\beta}). \tag{3.9}$$

Da der erste der beiden Summanden von $\boldsymbol{\beta}$ unabhängig ist, kann dieser bei der Maximierung der Loglikelihood Funktion vernachlässigt werden. Die Maximierung des verbleibenden Terms $-\frac{1}{2\sigma^2}\text{SSE}(\boldsymbol{\beta})$ entspricht der Minimierung von $\text{SSE}(\boldsymbol{\beta})$. Der Kleinste Quadrate Schätzer $\hat{\boldsymbol{\beta}}$ und der Maximum-Likelihood Schätzer $\hat{\boldsymbol{\beta}}$ stimmen demnach überein.

Schätzer der Fehlervarianz σ^2

Mittels Maximum-Likelihood Schätzung ergibt sich als Schätzer für die Varianz folgender Ausdruck, Fahrmeir, Kneib und Lang, 2009

$$\hat{\sigma}^2 = \frac{1}{n}\text{SSE}(\hat{\boldsymbol{\beta}}). \tag{3.10}$$

Da dieser nicht erwartungstreu ist, wird der erwartungstreue Schätzer verwendet

$$\tilde{\sigma}^2 = \frac{1}{n-p}\text{SSE}(\hat{\boldsymbol{\beta}}). \tag{3.11}$$

3.3.2 Streuungszerlegung und Bestimmtheitsmaß

In der Regressionsanalyse ist die Zerlegung der Totalen Quadratsumme

$$\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.12}$$

von zentraler Bedeutung. Basierend auf dieser Streuungszerlegungsformel kann das Bestimmtheitsmaß definiert werden, welches ein Maß für die Güte der Anpassung des Regressionsmodells darstellt.

Es gilt

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \bar{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ \text{SST} &= \text{SSR}(\hat{\boldsymbol{\beta}}) + \text{SSE}(\hat{\boldsymbol{\beta}}). \end{aligned} \tag{3.13}$$

Die Regressions-Quadratsumme $\text{SSR}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ beurteilt den Unterschied zwischen dem Regressionsmodell und einem Modell ohne Prädiktoren. Je größer die Regressions-Quadratsumme ist, desto aussagekräftiger ist das Modell.
Die Fehler-Quadratsumme $\text{SSE}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ wird durch den Kleinste Quadrate Schätzer minimiert.

Um die Güte eines Regressionsmodells zu bewerten, kann das Bestimmtheitsmaß verwendet werden. Es ist definiert als

$$\text{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{SSE}(\hat{\boldsymbol{\beta}})}{\text{SST}} = \frac{\text{SSR}(\hat{\boldsymbol{\beta}})}{\text{SST}}, \qquad 0 \leq \text{R}^2 \leq 1 \tag{3.14}$$

und gibt den durch das Regressionsmodell erklärten relativen Variationsanteil an, siehe Friedl, 2008.

Dafür gilt

- $\text{R}^2 = 1$: perfekte Anpassung des Modells ($\text{SSE}(\hat{\boldsymbol{\beta}}) = 0$) und
- $\text{R}^2 = 0$: keine lineare Abhängigkeit ($\text{SSR}(\hat{\boldsymbol{\beta}}) = 0$).

Je mehr Prädiktoren im Modell enthalten sind, desto größer wird das Bestimmtheitsmaß. Um diesem entgegenzuwirken, wird das Bestimmtheitsmaß korrigiert. Das adjustierte Bestimmtheitsmaß ist folgendermaßen definiert

$$R^2_{adj} = 1 - \frac{SSE(\hat{\boldsymbol{\beta}})/(n-p)}{SST/(n-1)}, \qquad 0 \leq R^2 \leq 1. \tag{3.15}$$

Bemerkung

- Ein anderes Maß zur Bewertung der Modellgüte ist durch $\hat{\sigma}$ gegeben. Je kleiner die Streuung der Residuen ist, umso besser passt das Modell. Im Gegensatz zum R^2, das keine Einheit hat, hängt $\hat{\sigma}$ von der Einheit der Zielvariablen ab, siehe Faraway, 2004.

3.3.3 Konfidenz- und Vorhersageintervall

In der Regression kann für eine neue gegebene Beobachtung $\mathbf{x}_+ = (1, x_{+,1}, \ldots, x_{+,p-1})$ der Wert für die Zielvariable vorhergesagt werden: $\hat{\mathbf{y}}_+ = \mathbf{x}_+^T \hat{\boldsymbol{\beta}}$. Wenn dieser Wert berechnet wurde, lassen sich die zugehörigen Konfidenzintervalle bzw. Vorhersageintervalle bestimmen, vgl. Fahrmeir, Kneib und Lang, 2009.

Das Konfidenzintervall gibt Auskunft über den Erwartungswert der Zielvariablen. Es überdeckt den Erwartungswert einer neuen Beobachtung $E(\mathbf{y}_+)$ an der Stelle $\mathbf{x}_+$ mit Wahrscheinlichkeit $(1-\alpha)$ und ist definiert als

$$\hat{\mathbf{y}}_+ \pm t_{n-p,1-\alpha/2} \sqrt{\hat{\sigma}^2 \mathbf{x}_+^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_+}. \tag{3.16}$$

Das Vorhersageintervall gibt Auskunft zu konkreten Einzelwerten. Es überdeckt eine neue Beobachtung $\mathbf{y}_+$ an der Stelle $\mathbf{x}_+$ mit Wahrscheinlichkeit $(1-\alpha)$ und ist definiert als

$$\hat{\mathbf{y}}_+ \pm t_{n-p,1-\alpha/2} \sqrt{\hat{\sigma}^2 (1 + \mathbf{x}_+^T (\mathbf{X}^T\mathbf{X})^{-1} \mathbf{x}_+)}. \tag{3.17}$$

Bemerkung

- Das Vorhersageintervall ist per Definition stets breiter als das zugehörige Konfidenzintervall.

3.4 Modelldiagnose

Die Schlussfolgerungen in der multiplen linearen Regressionsanalyse basieren auf einigen Annahmen. Falls diese nicht erfüllt sind, können Fehlschlüsse gezogen werden. Außerdem ist es wichtig, Ausreißer ausfindig zu machen, die das Modell unverhältnismäßig stark beeinflussen. Dieser Abschnitt gibt einen Überblick über einige

Diagnoseplots, die zur Überprüfung der Modellannahmen herangezogen werden können. Abbildung 3.6 bildet die in der Arbeit verwendeten Diagnoseplots ab.

Folgende Annahmen sind dabei zu überprüfen, siehe Pruscha, 2006

1. **Homoskedastizität**
   Die Varianzen σ^2 der Fehler ϵ_i sind für alle Beobachtungen $i = 1, \ldots, n$ identisch.
2. **Linearität**
   Der Erwartungswert der Zielvariablen hängt linear von den Prädiktoren ab.

$$\mathrm{E}(y_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_{p-1} x_{i,p-1}$$

3. **Unabhängigkeit**
   Die Fehler $\epsilon_i,\ i = 1, \ldots, n$, sind stochastisch unabhängig.
4. **Normalverteilung**
   Die Fehler $\epsilon_i,\ i = 1, \ldots, n$, sind normalverteilt.

Die Voraussetzungen (1), (3) und (4) sind unter Zuhilfenahme der Fehler ϵ_i notiert, da die Residuen e_i zur Prüfung der Annahmen eingesetzt werden. Dabei sind die Residuen $e_i = y_i - \hat{y}_i$ die Schätzer der nicht beobachtbaren Zufallsfehler ϵ_i. Um die Anpassung des linearen Modells und die entsprechenden Modellannahmen zu überprüfen, eignen sich besonders visuelle Methoden. Zur Überprüfung der Annahmen werden die folgenden Diagnoseplots verwendet.

Scatterplot: Vorhersagewerte gegen Residuen

Der Scatterplot der Vorhersagewerte gegen die Residuen stellt im Idealfall eine horizontal ausgerichtete Punktwolke dar. Abweichungen davon deuten auf Verletzungen der Varianzhomogenität und Linearität hin.

Q-Q-Plot der Residuen

Mit Hilfe des Q-Q-Plots der Residuen wird die Normalverteilungsannahme der Zufallsfehler grafisch getestet. Falls eine Normalverteilung vorliegt, liegen die Punkte auf der Referenzlinie. Abweichungen vom linearen Muster weisen auf eine Verletzung der Normalverteilungsannahme hin.

Scale-Location Plot

Die Annahme der Homoskedastizität kann durch einen Scatterplot der Vorhersagewerte gegen die Wurzeln aus dem Betrag der standardisierten Residuen überprüft werden. Die standardisierten Residuen

$$e_i^* = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_{ii}}} \tag{3.18}$$

sind dazu besser geeignet als die gewöhnlichen Residuen, vgl. Fahrmeir, Kneib und Lang, 2009. Bei Homoskedastizität sollten die aufgetragenen Punkte eine zufällige

Struktur ohne Muster bilden. Systematische Trends weisen auf eine Varianzinhomogenität hin. Falls die Annahme der konstanten Varianzen verletzt ist, kann das Problem mittels varianzstabilisierenden Transformationen behoben werden, vgl. Abschnitt 3.5.

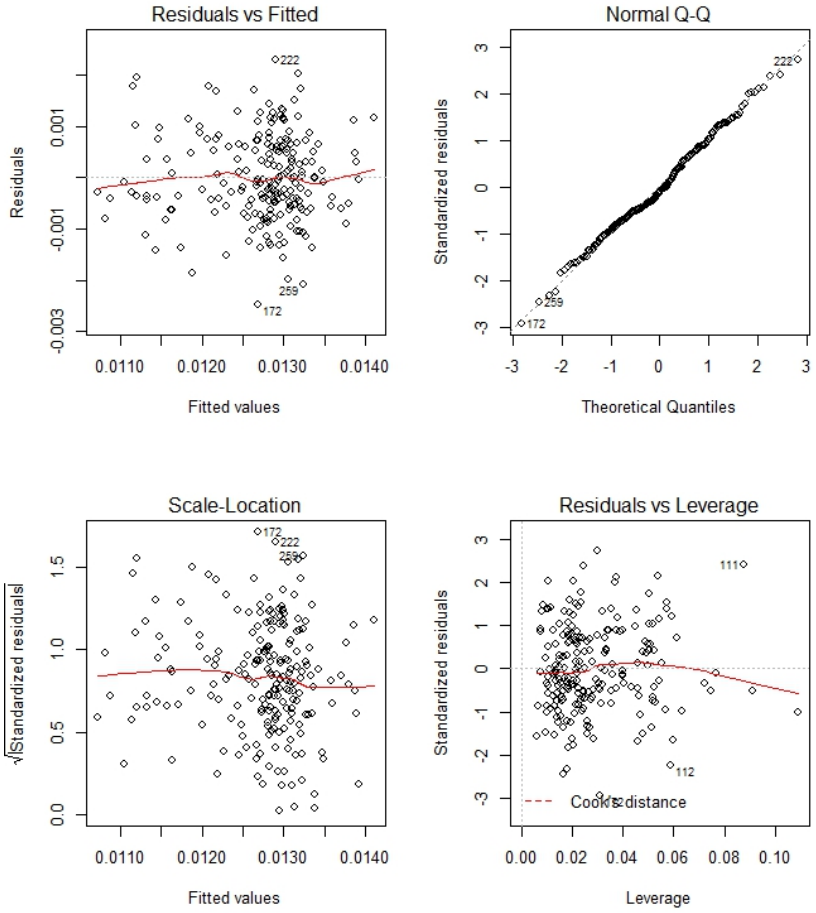

Abbildung 3.6: Modelldiagnoseplots des linearen Regressionsmodells für die mittlere Goldschichtstärke der kleinen Pads im reduzierten Modellierungsdatensatz (ENIG.Reg).

Scatterplot: Hebelwerte gegen Vorhersagewerte

Beobachtungen mit einem großen Einfluss können im Scatterplot der Hebelwerte (Diagonalelemente der Hat-Matrix (Leverages)) gegen Vorhersagewerte gefunden werden. Die Hebelwerte bewegen sich im Intervall $[1/n; 1]$. Ein großer Hebelwert bedeutet, dass die zugehörige Beobachtung einen großen Einfluss auf die Schätzung hat.

Bemerkung

- Laut Fahrmeir, Kneib und Lang, 2009, sollten Beobachtungen mit $h_{ii} > 2p/n$ genauer untersucht werden. Dazu ist zu bemerken, dass große Hebelwerte nicht

immer auf Probleme hinweisen müssen. Als Kriterium für den Einfluss von Hebelwerten auf die Parameterschätzung wird die Cook-Distanz verwendet. Dabei sind Werte mit Cook-Distanzen größer als 0.5 genauer zu analysieren.

3.5 Transformationen

Die Regressionsanalyse geht von der Annahme der Linearität und der Homoskedastizität der zufälligen Fehler aus. Falls diese Annahmen nicht erfüllt sind, besteht die Möglichkeit, mit Hilfe geeigneter Transformationen das Problem nicht-konstanter Fehlervarianzen zu beheben, vgl. Friedl, 2014.

Die Box-Cox Transformation bietet die Möglichkeit, eine passende Transformation zu bestimmen. Das Ziel dabei ist, einen Parameter λ für eine varianzstabilisierende Transformation zu finden, sodass die Varianz der Zielvariablen $y^\star$ unabhängig von deren Erwartungswert ist. Die Transformationsfunktion ist gegeben durch

$$y^\star(\lambda) = \begin{cases} \dfrac{y^\lambda - 1}{\lambda}, & \text{falls } \lambda \neq 0, \\ \log y, & \text{falls } \lambda = 0. \end{cases} \tag{3.19}$$

Zur Bestimmung des perfekten λ wird die Profile Likelihood Funktion maximiert. Nähere Informationen dazu finden sich in Friedl, 2014.

In **R** kann zur Berechnung des Schätzers von λ die Funktion *boxcox* des Pakets *MASS* verwendet werden, siehe Ripley u. a., 2015. Diese zeichnet die Profile Likelihood Funktion inklusive eines 95% Konfidenzintervall für λ, siehe Abbildung 3.7. In der Praxis wird für die Transformation eine Zahl innerhalb des Konfidenzintervalls gewählt; in diesem Fall wäre ein Wahl $\hat{\lambda} = -1$ geeignet.

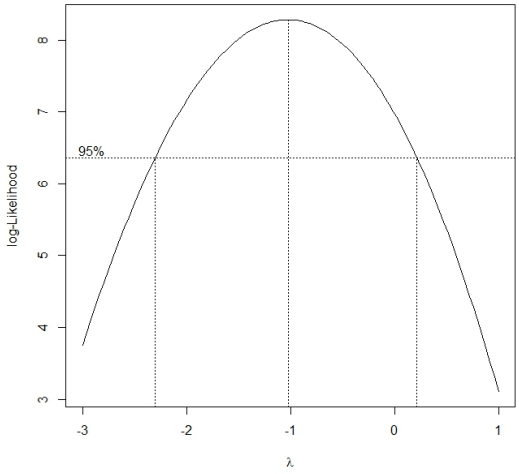

Abbildung 3.7: Box-Cox Transformationsplot. Profile Likelihood Funktion mit 95% Konfidenzintervall für λ.

Bemerkung

- Transformationen finden auch andere Anwendungsgebiete. Zum Beispiel eignet sich die lineare Regression auch zur Beschreibung nicht-linearer Zusammenhänge, wie in Burkschat, Cramer und Kamps, 2012, erläutert ist. Durch eine geeignete Transformation können in manchen Fällen nicht-lineare Zusammenhänge linearisiert werden, zb. $Y = ae^{bX} \Rightarrow \ln(Y) = \ln(a) + bX$.
- Ebenso können nicht-lineare Einflüsse der Prädiktoren durch Variablentransformationen modelliert werden. Dabei führen unterschiedliche Transformationen oft zu ähnlichen Schätzungen. Da es keine Patentlösung zur Bestimmung vernünftiger Variablentransformationenen gibt, ist die Überprüfung der Scatterplots und Residuen notwendig, siehe Fahrmeir, Kneib und Lang, 2009.

3.6 Modellbildung

Mit Hilfe der linearen Regressionsanalyse können Zusammenhänge zwischen zwei oder mehreren Variablen untersucht werden. Stehen mehrere Prädiktoren zu Verfügung, stellt sich die Frage, welche Variablen in das Modell aufgenommen werden sollen. Für m potentielle Prädiktoren ergeben sich 2^m mögliche Regressionsmodelle, falls alle Teilmengen betrachtet werden. Ziel ist, das beste der 2^m Modelle auszuwählen, vgl. Friedl, 2008.

3.6.1 Modellwahlkriterien

Um Modelle auszuwählen und zu beurteilen stehen verschiedene Kriterien zur Verfügung.

Bestimmtheitsmaß R^2_{adj}

Wie schon beschrieben wird das adjustierte Bestimmtheitsmaß zur Beurteilung der Güte der Modellanpassung verwendet (siehe Gleichung (3.14))

$$R^2 = \frac{SSR(\hat{\boldsymbol{\beta}})}{SST} = 1 - \frac{SSE(\hat{\boldsymbol{\beta}})}{SST}.$$

Da das Bestimmtheitsmaß R^2 mit der Anzahl an Prädiktoren (auch irrelevanten) größer wird, kann das adjustierte Bestimmtheitsmaß R^2_{adj} (3.15) berechnet werden

$$R^2_{adj} = 1 - \frac{SSE(\hat{\boldsymbol{\beta}})/(n-p)}{SST/(n-1)},$$

wobei p gleich die Anzahl an Prädiktoren inklusive Intercept im Modell darstellt.

Bemerkung

- $R^2 \approx 0$ bedeutet nicht, dass kein Zusammenhang zwischen der Zielvariable und den Prädiktoren besteht. Nichtlineare Zusammenhänge können durchaus bestehen, siehe Friedl, 2008.

Mallows´ C_p - Statistik

Mallows´ C_p Kriterium basiert auf dem vollständigen Modell und bewertet Teilmodelle in Bezug auf ihre Prognosequalität, siehe Gross, 2010. Die Statistik von Mallows ist definiert als

$$C_p = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_{i(p)})^2}{\hat{\sigma}^2} - n + 2p = \frac{\mathrm{SSE}(\hat{\boldsymbol{\beta}}_p)}{\mathrm{SSE}(\hat{\boldsymbol{\beta}}_m)/(n-m)} - n + 2p, \tag{3.20}$$

wobei $\hat{\sigma}^2$ der quadrierte Residuenstandardfehler des vollständigen Modells ist, $\boldsymbol{\beta}_p$ der Koeffizientenvektor des betrachteten Modells ($p < m$) und $\boldsymbol{\beta}_m$ der Koeffizientenvektor des vollständigen Modells.

Für das vollständige Modell gilt

$$C_p = \frac{\mathrm{SSE}(\hat{\boldsymbol{\beta}}_m)}{\mathrm{SSE}(\hat{\boldsymbol{\beta}}_m)/(n-m)} - n + 2m = m. \tag{3.21}$$

Je kleiner der C_p-Wert ist, umso besser ist das Modell.

Akaike´s Informationskriterium (AIC)

Das Informationskriterium bewertet die Anpassungsgüte des Regressionsmodells durch die negative Likelihood des Modells und ist definiert als

$$\mathrm{AIC} = -2\,\mathrm{logL}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{y}) + 2(p+1). \tag{3.22}$$

Das Modell ist umso besser, je kleiner der Wert des Akaike Informationskriteriums ist.

Korrigiertes Akaike Informationskriterium (AIC_c)

Die korrigierte Fassung des Informationskriteriums ist definiert als

$$\begin{aligned} \mathrm{AIC}_c &= -2\,\mathrm{logL}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{y}) + 2(p+1) + 2\frac{(p+1)(p+2)}{n-p} \\ &= \mathrm{AIC} + 2\frac{(p+1)(p+2)}{n-p}. \end{aligned} \tag{3.23}$$

Diese Version eignet sich besonders für kleine Stichprobengrößen bzw. für eine im Vergleich zum Stichprobenumfang große Anzahl an Prädiktoren ($n/(p+1) \leq 40$).

Bayes´sches Informationskriterium (BIC)

Das Informationskriterium ist derart definiert, dass das beste Modell jenes mit dem kleinsten BIC ist

$$\text{BIC} = -2\,\text{logL}(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2, \mathbf{y}) + (p+1)\log(n). \tag{3.24}$$

Der Unterschied zwischen den Informationskriterien AIC und BIC ist sehr gering. Statt im Strafterm den Faktor 2 zu verwenden, benützt das BIC den Faktor $\log(n)$.

Bemerkung

- Durch die Bewertung mittels BIC werden Modelle mit weniger Prädiktoren bevorzugt, siehe Fahrmeir, Kneib und Lang, 2009.

3.6.2 Praktische Anwendung der Modellwahlkriterien

Um aus einer Sammlung an Modellen das vielversprechendste Modell auszuwählen, können die vorgestellten Modellwahlkriterien verwendet werden. Fahrmeir, Kneib und Lang, 2009, empfiehlt folgende Vorgehensweise:

- Anhand verschiedener Vorüberlegungen bzw. durch Vorwissen wird eine geringe Anzahl möglicher Modelle ausgewählt. Die ausgewählten Modelle können sich durchaus in der Anzahl der Prädiktoren unterscheiden. Auch die Art der Modellierung (linear, nicht-linear) muss nicht dieselbe sein.
- Die ausgewählten Modelle können anschließend mit den genannten Modellwahlkriterien beurteilt werden. Dabei sollte nicht ausschließlich das beste Modell gewählt werden, da meistens mehrere Modelle annähernd gleich bewertet werden und sich nur geringfügig unterscheiden.

Diese Methode ist leider nicht immer zweckmäßig, da vielmals die Anzahl der zur Verfügung stehenden Prädiktoren und Modellierungsvarianten sehr groß ist. So ist die Berechnung aller möglichen Modelle nicht einfach durchführbar. In diesem Zusammenhang können folgende Verfahren angewendet werden:

Vollständige Modellselektion (All-Subset-Selection)

Falls die Anzahl der zur Verfügung stehenden Prädiktoren kleiner als 40 ist, kann das entsprechend der Modellwahlkriterien beste Modell eruiert werden. Im zugehörigen Algorithmus werden nicht alle Modelle berechnet. Im **R**-Paket *leaps* findet sich eine Implementierung. Für nähere Informationen wird auf Lumley, 2015, verwiesen.

Vorwärts-Selektion (Forward-Selection)

Ausgehend von einem Startmodell, das nur den Intercept enthält, wird in jeder Iteration ein Prädiktor aus dem Variablenpool in das Modell inkludiert, sodass das resultierende Modell das Modellwahlkriterium so gut wie möglich verbessert. Der Algorithmus endet, wenn alle Prädiktoren im Modell enthalten sind oder keine weitere Verbesserung des Modellwahlkriteriums erreichbar ist.

Rückwärts-Selektion (Backward-Selection)

Die Rückwärts-Selektion ist das Gegenteil der Vorwärts-Selektion. Diese startet mit der Modellierung eines Regressionsmodells, in dem alle verfügbaren Prädiktoren enthalten sind. Anschließend wird in jeder Iteration jener Prädiktor entfernt, der das Modellwahlkriterium bestmöglich verbessert. Der Algorithmus endet, wenn alle Prädiktoren entfernt wurden oder keine weitere Verbesserung des Modellwahlkriteriums realisierbar ist.

Bemerkung

- Sowohl die Rückwärts-Selektion als auch die Vorwärts-Selektion betrachten maximal $m(m+1)/2$ mögliche Teilgruppen von Prädiktoren, siehe Friedl, 2008.
- In jeder Iteration wird ein Modell mit dem kleinsten Wert des Informationskriteriums gewählt. Dies entspricht den Prädiktor mit dem größten p-Wert zu entfernen, siehe Friedl, 2008.

Schrittweise-Selektion (Stepwise-Selection)

Die schrittweise Regression vereint die Vorwärts- und Rückwärts-Selektionen. In jeder Iteration wird entschieden, ob eine Variable in das Modell aufgenommen oder entfernt wird.

Die Verfahren Vorwärts-, Rückwärts- und Schrittweise-Selektion liefern grundsätzlich sehr gute Modelle, aber nicht das beste Modell hinsichtlich der Modellwahlkriterien.

Bemerkung

- Die Behandlung hierarchischer Terme kann ein Problem darstellen. Werden etwa Polynome verwendet, um nicht-lineare Effekte zu modellieren, so können die Algorithmen quadratische und kubische Anteile in das Modell aufnehmen, jedoch den linearen Anteil verwerfen, vgl. Fahrmeir, Kneib und Lang, 2009.

3.7 Regressionsbäume

Eine Möglichkeit, Regressionsmodelle grafisch darzustellen, bilden Regressionsbäume. Dabei werden Daten anhand einer quantitativen Zielvariablen nach speziellen Charakteristiken in Gruppen aufgeteilt, vgl. Bankhofer und Vogel, 2008.

Regressionsbäume sind mit einem Wurzel(knoten), inneren Knoten, Kanten und Blattknoten „baumähnlich“ aufgebaut. Dabei entsprechen die Wurzel und inneren Knoten Merkmalen, nach denen die Daten aufgeteilt werden und die Blätter stellen die zugehörigen Daten in Form von z.B. Boxplots dar, sowie die in dieser Arbeit verwendete **R**-Funktion *ctree* aus dem Paket *party*, siehe Hothorn u. a., 2015.

Ein Regressionsbaum, siehe Abbildung 3.8, entsteht durch rekursives Partitionieren der Beobachtungen, siehe Grömping, 2009. Das erste Merkmal, welches die Aufteilung der Daten in zwei Gruppen startet, bildet den Wurzelknoten. Anschließend werden die Teildatensätze in mehr und mehr homogene Gruppen aufgeteilt, welche die einzelnen Knoten bzw. Blätter bilden. Jeder Split basiert auf den Werten einer Variablen und wird nach speziellen Kriterien ausgewählt.

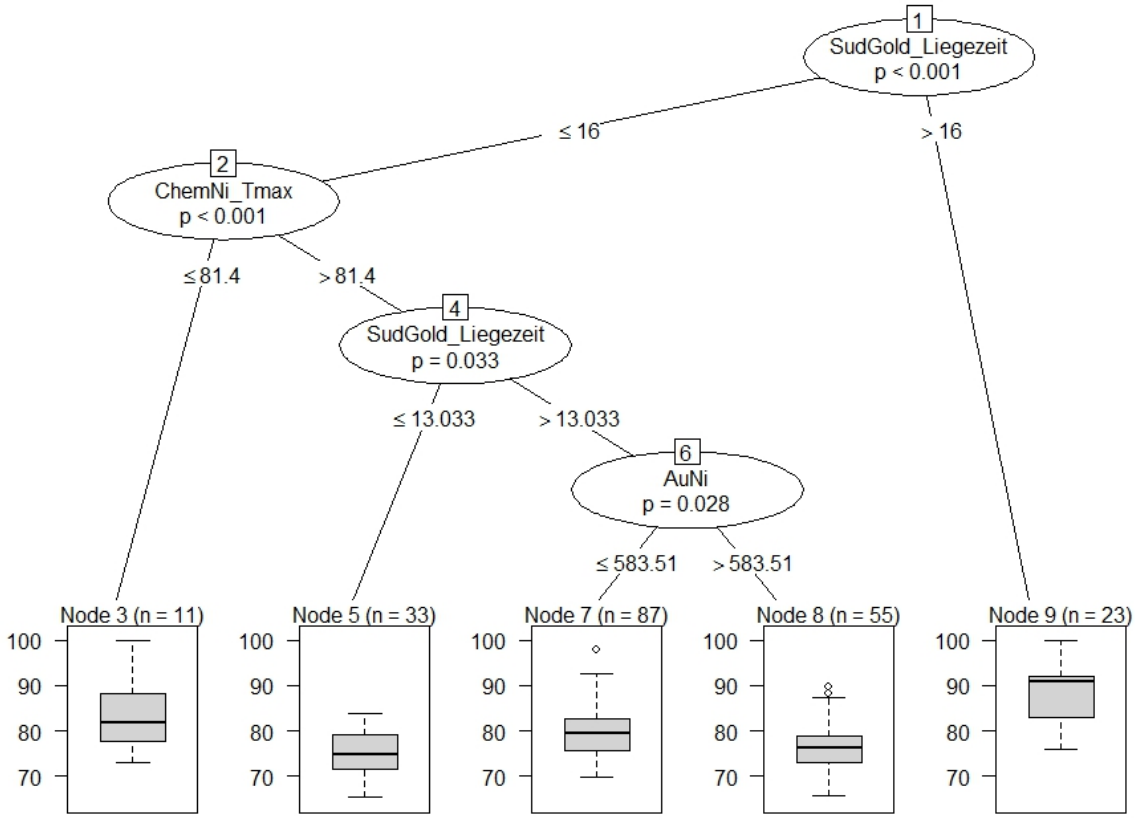

Abbildung 3.8: Regressionsbaum des linearen Regressionsmodells für die mittlere Goldschichtstärke der kleinen Pads im reduzierten Modellierungsdatensatz (ENIG.Reg). Die beobachteten Daten werden anhand der Prädiktoren des Regressionsmodells in Gruppen aufgeteilt.

Sobald ein Baum gebildet wurde, kann die Zielvariable für jede Beobachtung vorhergesagt werden, indem dem Pfad vom Wurzelknoten bis zum entsprechenden Blattknoten gefolgt wird. Der Wert der vorhergesagten Zielvariable ist die mittlere Zielvariable des Blattknotens.

3.8 Multivariate multiple lineare Regression

In diesem Abschnitt wird die Theorie der multivariaten linearen Regressionsmodelle, wie in Fox und Weisberg, 2011 und Maitra, 2013 ausgeführt, kurz beschrieben.

Die multivariate multiple lineare Regressionsanalyse basiert auf der Modellierung zweier oder mehrerer Zielvariablen. Dabei werden alle Zielvariablen $\mathbf{y}_1, \ldots, \mathbf{y}_m$ mit dem gleichen Satz an Prädiktoren $\mathbf{x}_1, \ldots, \mathbf{x}_{p-1}$ modelliert. Für jede Zielvariable ergibt sich ein eigenes Regressionsmodell

$$\begin{aligned}
\mathbf{y}_1 &= \beta_{01} + \beta_{11}\mathbf{x}_1 + \beta_{21}\mathbf{x}_2 + \cdots + \beta_{p-1,1}\mathbf{x}_{p-1} + \boldsymbol{\epsilon}_1, \\
\mathbf{y}_2 &= \beta_{02} + \beta_{12}\mathbf{x}_1 + \beta_{22}\mathbf{x}_2 + \cdots + \beta_{p-1,2}\mathbf{x}_{p-1} + \boldsymbol{\epsilon}_2, \\
&\vdots \\
\mathbf{y}_m &= \beta_{0m} + \beta_{1m}\mathbf{x}_1 + \beta_{2m}\mathbf{x}_2 + \cdots + \beta_{p-1,m}\mathbf{x}_{p-1} + \boldsymbol{\epsilon}_m.
\end{aligned} \tag{3.25}$$

Das multivariate lineare Regressionsmodell fasst die Modelle zusammen und ist definiert als

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}, \tag{3.26}$$

wobei die $n \times m$ Matrix $\mathbf{Y}$ die n Beobachtungen der m Zielvariablen beschreibt,

$$\mathbf{Y}_{n\times m} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{pmatrix},$$

die $n \times p$ Designmatrix $\mathbf{X}$ Spalten für den Intercept und alle $p-1$ Prädiktoren beinhaltet,

$$\mathbf{X}_{n\times p} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{pmatrix},$$

die $p \times m$ Matrix $\mathbf{B}$ die Regressionskoeffizienten zusammenfasst (eine Spalte pro Zielvariable),

$$\mathbf{B}_{p\times m} = \begin{pmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{p-1,1} & \beta_{p-1,2} & \cdots & \beta_{p-1,m} \end{pmatrix}$$

und $\mathbf{E}$ die $n \times m$ Fehlermatrix ist,

$$\mathbf{E}_{n\times m} = \begin{pmatrix} \epsilon_{11} & \epsilon_{12} & \cdots & \epsilon_{1m} \\ \epsilon_{21} & \epsilon_{22} & \cdots & \epsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \epsilon_{n1} & \epsilon_{n2} & \cdots & \epsilon_{nm} \end{pmatrix}.$$

Bemerkung

- Die Designmatrix $\mathbf{X}$ des multivariaten Modells unterscheidet sich von der Designmatrix des univariaten Modells nicht.

Wie im univariaten linearen Modell werden folgende Modellannahmen getroffen

- $\boldsymbol{\epsilon}_i^T \sim \text{Normal}_m(\mathbf{0}, \boldsymbol{\Sigma})$, wobei $\boldsymbol{\epsilon}_i^T$ die i-te Zeile der Fehlermatrix $\mathbf{E}$ ist und $\boldsymbol{\Sigma}$ eine reguläre Fehler-Kovarianzmatrix (konstant über alle Beobachtungen) darstellt,
- $\boldsymbol{\epsilon}_i^T$ und $\boldsymbol{\epsilon}_j^T$ sind unabhängig für $i \neq j$ und
- $x_{ij},\ i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, p-1$ sind bekannte Konstanten.

Um die $p \times m$ Matrix $\mathbf{B}$ der Regressionskoeffizienten des multivariaten Regressionsmodells zu schätzen, müssen die gegebenen Beobachtungen herangezogen werden.

Der Maximum-Likelihood Schätzer von $\mathbf{B}$ im multivariaten linearen Modell ist gegeben durch

$$\hat{\mathbf{B}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \tag{3.27}$$

und ist äquivalent zum Kleinste Quadrate Schätzer der einzelnen Zielvariablen.

4 Praktische Problemlösung

Ein zentraler Prozess in der Leiterplattenherstellung ist der ENIG-Prozess (siehe Kapitel 2.2, Der ENIG-Prozess). Im Zuge dieses Prozesses wird eine funktionale Schutzschicht auf die Kontaktstellen einer Leiterplatte aufgebracht um diese unter anderem vor Oxidation zu schützen. Die Schutzschicht besteht aus einer Nickelschicht und einer Goldschicht.

Ziel dieser Arbeit ist die statistische Analyse der Prozessparameter des ENIG-Prozesses und deren Zusammenhang mit der Schichtstärke der Schutzschicht, insbesondere der Goldschichtstärke. Prozessparameter mit einem signifikanten Einfluss auf die Goldschichtstärke werden mit Hilfe von explorativen Analysen konkretisiert. Anschließend wird die Goldschichtstärke mittels multivariater Regressionsmodelle modelliert. Damit soll eine Optimierung der Prozessführung erreicht werden, indem der Verbrauch des Edelmetalls minimiert wird.

Dieses Kapitel gliedert sich in drei Teile. Die Vorstudie, *Abschnitt 4.1, Historische Daten*, basiert auf einem Datensatz, der seit 2012 erfasst wurde. Diese Daten geben einen ersten Einblick in die vorliegende Problematik. Die statistische Analyse ergab eine ausbaufähige Datenqualität des Datensatzes, schwache Zusammenhänge der Prozessparameter und eine damit verbundene geringe Modellgüte des modellierten Regressionsmodells. Zum Zweck der Verbesserung der Schwachpunkte des historischen Datensatzes wurde ein neuer modifizierter Datensatz erstellt. Dabei wurde die Messprozedur adaptiert und weitere Prozessparameter in den Datensatz hinzugenommen. Die Analyse und Modellierung dieses neuen Datensatzes bilden *Abschnitt 4.4, Modellierungsdaten*. Die statistischen Analysen erzielten mit dem neuen Datensatz ein ähnliches Ergebnis: ausbaufähige Datenqualität, schwache Korrelationen und Regressionsmodelle mit einer geringen Anpassungsgüte. Zum Schluss wurde das entwickelte Regressionsmodell anhand eines weiteren Datensatzes validiert. *Abschnitt 4.7, Validierung der Regressionsmodelle*, enthält die entsprechenden Auswertungen.

4.1 Historische Daten

Für die erste statistische Untersuchung der Daten des ENIG-Prozesses wurde ein vorhandener Datensatz herangezogen. Dafür wurden Testleiterplatten, siehe Abbildung 4.1, produziert und in der laufende Produktion im ENIG-Prozess mitgeführt. Einige Prozessparameter und die zugehörigen Nickel- und Goldschichtstärken der Testleiterplatten wurden in einem Datensatz festgehalten.

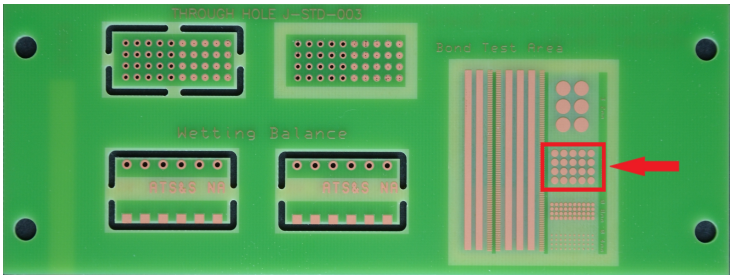

Abbildung 4.1: Testleiterplatte und Messprozedur der Schichtstärke. Die Nickel- bzw. Goldschichtstärke wurde auf acht verschiedenen (nicht genau spezifizierten) Pads mittlerer Größe gemessen.

4.1.1 Beschreibung des Datensatzes

Der Datensatz beinhaltet 1055 Beobachtungen, die im Zeitraum vom 15.02.2012 bis 06.11.2014 gemessen wurden.

Die Parameter des Datensatzes sind in Tabelle 4.1 aufgelistet. Um eine Beobachtung eindeutig identifizieren zu können, wurde sie mit einer Nummer beschriftet. Die zeitliche Einordnung erfolgt mit Hilfe der Parameter Datum und Schicht. Als Prozessparameter der Aktivstationen des ENIG-Prozesses wurden jeweils die gleichen vier Parameter im Nickelbad als auch im Goldbad gemessen. Diese bestehen aus dem pH-Wert, der Solltemperatur, der Konzentration und des Metal-Turn-Overs (MTO) beider Bäder.

Tabelle 4.1: Variablen im historischen Datensatz (ENIG.alt). Neben dem Zeitstempel sind pro Beobachtung acht Prozessparameter, die Nickelschichtstärken und die Goldschichtstärken festgehalten.

	Variable	**Abkürzung**	**Einheit**
Zeitstempel	Nummer	Nummer	
	Datum	Datum	
	Schicht	Schicht	
Nickelbad	pH-Wert	NipH	
	Solltemperatur [1]	NiTemp	$[°C]$
	Konzentration [2]	NiKonz	$[g/l]$
	MTO [3]	NiMTO	
Goldbad	pH-Wert	AupH	
	Solltemperatur [1]	AuTemp	$[°C]$

Fortsetzung nächste Seite

[1]Die Solltemperatur ist die Einstelltemperatur der jeweiligen Aktivstation, Breitwieser, 2015.

[2]Die Konzentration gibt Auskunft über die Anteile eines bestimmten Stoffes (hier: Nickel bzw. Goldsalz) im gesamten Badvolumen, Breitwieser, 2015.

[3]Der MTO (Metal-Turn-Over) gibt das Alter des Bades an. Er erhöht sich bei jeder Zugabe von Nickel bzw. Goldsalz. Da die Bäder keine unendliche Lebensdauer besitzen, müssen sie regelmäßig erneuert werden. Der MTO wird dabei wieder auf Null gesetzt, Breitwieser, 2015.

Tabelle 4.1 – *Fortsetzung von vorheriger Seite*

	Variable	**Abkürzung**	**Einheit**
	Konzentration [2]	AuKonz	$[g/l]$
	MTO [3]	AuMTO	
Schichtstärken	Nickelschichtstärke: Wert 1 bis Wert 8	Ni1 bis Ni8	$[\mu m]$
	Goldschichtstärke: Wert 1 bis Wert 8	Au1 bis Au8	$[\mu m]$
	Mittelwert der Nickelschichtstärken	MWNi	$[\mu m]$
	Mittelwert der Goldschichtstärken	MWAu	$[nm]$

Bemerkung

- Die dritte Spalte in Tabelle 4.1 gibt die Abkürzungen der Variablen an, die in den nachfolgenden Grafiken verwendet werden.

Die Schichtstärken wurden auf acht unterschiedlichen Pads mittlerer Größe ermittelt. Diese haben einen Durchmesser von 1.6 mm. Um einen stabilen, repräsentativen Wert der Schichtstärken für die Auswertungen zu bekommen, wurde für die Datenauswertung der Mittelwert der Messungen der Schichtstärken pro Beobachtung berechnet. So entstand ein gemittelter Wert der Nickelschichtstärke und ein gemittelter Wert der Goldschichtstärke pro Beobachtung.

Ebenso wurden der Median und die Standardabweichung der Messungen der Schichtstärken der einzelnen Beobachtungen berechnet. Da der Unterschied zwischen Medianen und Mittelwerten sehr gering war und die Betrachtung der Standardabweichungen bis auf einige Ausreißer ein homogenes Bild ergab, wurden in weiterer Folge nur die Mittelwerte der Schichtstärken für die Analysen herangezogen.

Die entsprechenden Einheiten, in denen die Parameter gemessen wurden, sind ebenfalls in Tabelle 4.1 enthalten. Die Schichtstärken wurden in Mikrometer gemessen. Da die Goldschicht sehr dünn ist, wurde der betrachtete Mittelwert der Goldschichtstärke in Nanometer umgerechnet. Die Abweichungen der Schichtstärken bewegen sich folglich in einem sehr kleinen Bereich.

Aufschluss über die Messstellen im Prozessverlauf gibt Abbildung 4.2. Lediglich in zwei Aktivstationen wurden Parameter beobachtet. Einflüsse der anderen acht Aktivstationen wurden nicht berücksichtigt.

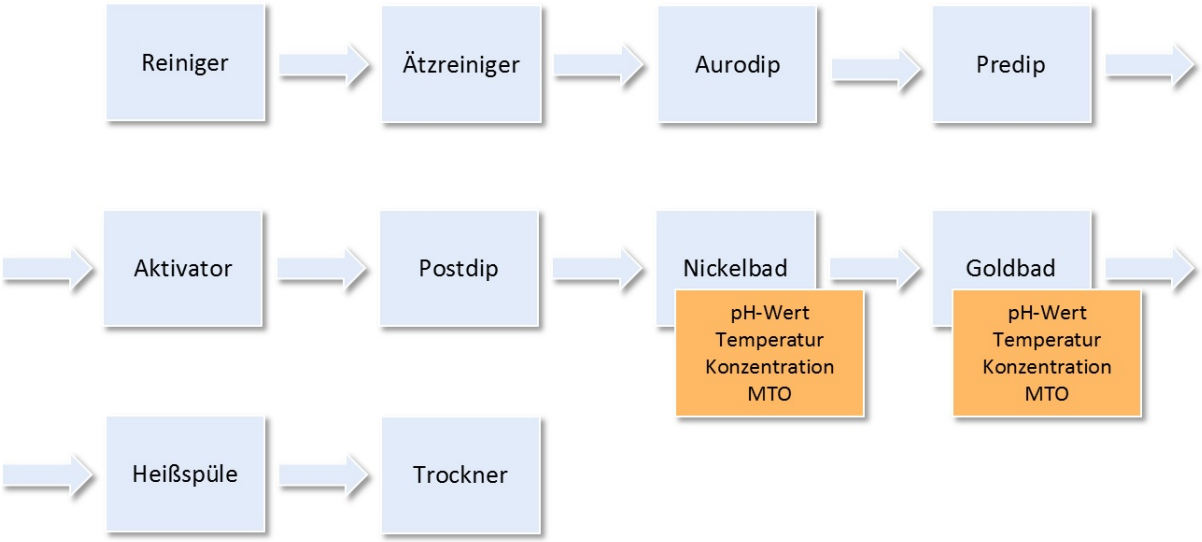

Abbildung 4.2: Variablen im Prozessverlauf im historischen Datensatz (ENIG.alt). Jeweils vier Prozessparameter der Aktivstationen Nickelbad und Goldbad wurden gemessen.

Zusätzlich wurden im beobachteten Zeitraum unterschiedliche Bäder verwendet, jeweils zwei verschiedene Nickelbäder mit der Abkürzungen CNN und NIC und zwei verschiedene Goldbäder mit den Abkürzungen ATS und AU. Dabei hat jedes Bad unterschiedliche chemische Zusammensetzungen. Aufgrund dessen handelt es sich um unterschiedliche Reaktionstypen, die nicht miteinander verglichen werden können. Weiters sind je nach Bad-Typ andere fixe Parametereinstellungen bzw. fixe Parametereinstellungsintervalle verwendet worden. Eine Analyse der Daten ohne Berücksichtigung der Bad-Zusammensetzungen ist daher nicht sinnvoll. Tabelle 4.2 stellt eine Übersicht dar, welche Bäder in welchem Zeitraum verwendet wurden. Dabei stellte sich heraus, dass alle Kombinationen bis auf CNN/AU auftreten.

Tabelle 4.2: Unterschiedliche Bad-Typen im historischen Datensatz (ENIG.alt). Während des Beobachtungszeitraums wurden verschiedene Typen von Nickel- und Goldbädern verwendet.

Nickelbad	CNN	Nummer 1 bis Nummer 160
	NIC	Nummer 161 bis Nummer 1055
Goldbad	ATS	Nummer 1 bis Nummer 173
		Nummer 325 bis Nummer 1055
	AU	Nummer 174 bis Nummer 324

4.1.2 Datenqualität und Datenbereinigung

„Datenqualität ist nicht alles, aber ohne Qualität der Daten ist alles nichts.“
(Schendera, 2007)

Dieses Kapitel folgt den Ausführungen in Schendera, 2007 und Naumann, 2007.

Die Qualität der Daten spielt in statistischen Auswertungen eine große Rolle. Statistische Studien liefern nur mit Daten, die eine entsprechende Qualität aufweisen, aussagekräftige Ergebnisse. Fehlerhafte Daten verursachen fehlerhafte statistische Auswertungen und in weiterer Folge Fehlentscheidungen. Demnach sollte auf eine hohe Datenqualität großer Wert gelegt werden.

Folgende primäre Qualitätsmerkmale sollte ein Datensatz aufweisen:

- Vollständigkeit
- kontrollierte Missings
- Vermeidung von doppelten Daten
- Einheitlichkeit
- Beurteilung von Ausreißern
- Plausibilität

Im ersten Schritt sollten die Kriterien Vollständigkeit, Missings und doppelte Daten überprüft werden. Anschließend werden Einheitlichkeit und Ausreißer beurteilt. Als letztes Kriterium sollte die Plausibilität herangezogen werden.

Das erste Kriterium ist die Vollständigkeit. Datenzeilen und Datenspalten müssen genau dem entsprechen, was beobachtet wurde. Zudem dürfen keine Daten willkürlich entfernt oder zugefügt werden. Fehlende Werte in einem Datensatz werden als Missings bezeichnet. Die Qualität statistischer Ergebnisse steht in direktem Zusammenhang mit dem Ausmaß von Missings. Als Daumenregel gelten Anteile von Missings bis zu 5% als zumutbar. Bei vollständigen Daten folgt die Überprüfung auf doppelte Daten. Diese verfälschen die Ergebnisse dahingehend, dass sie die Gewichtung verändern. Weiters muss die Einheitlichkeit auf verschiedenen Stufen sicher gestellt werden. Dazu zählen einheitliche Bezeichnungen der Variablen bei Verwendung mehrerer Datensätze, sowie einheitliche Bezeichnung von Ausprägungen für Variablen und einheitliche Datums-Variablen. Ausreißer stellen ein Problem für statistische Analysen dar. Sie können unter anderem Ergebnisse in der linearen Regression völlig verzerren. Abschließendes Kriterium ist die Plausibilität. Dabei werden Daten auf ihre inhaltliche Korrektheit überprüft.

Weitere Qualitätsmerkmale sind Menge, Eindeutigkeit, Relevanz, Genauigkeit, Verständlichkeit, Aktualität, Dokumentation, Glaubwürdigkeit, Objektivität, Status, Überprüfbarkeit, Kompatibilität, Verfügbarkeit, Ressourcenverbrauch und Preis/Kosten. Eine Beschreibung dieser Merkmale findet sich im Buch von Schendera, 2007.

Der historische Datensatz wurde zur Überprüfung und zu Dokumentationszwecken aufgezeichnet. Eine statistische Analyse der Daten, für welche die Datenqualität eine wesentliche Rolle spielt, war von Beginn an nicht geplant. Dementsprechend wurde kein Wert auf die Datenqualität gelegt. So waren im historischen Datensatz einige Qualitätsmerkmale nicht erfüllt. Nach Bereinigung der Daten - fehlende Werte, Schreibfehler, falsche Referenzen, unterschiedliche Labels, widersprüchliche Werte, unterschiedliche Einheiten, unterschiedliche Genauigkeiten, Duplikate, strukturelle Inhomogenität - blieben 758 vollständige Beobachtungen zur Analyse übrig. Knapp 72% der Daten konnten demnach verwendet werden.

Abbildung 4.3 stellt die zwei Datensätze, die für die Analyse verwendet wurden, dar. Der Datensatz ENIG.alt beinhaltet alle 758 ausgebesserten Beobachtungen. ENIG.NIC.ATS ist ein Teilmenge des Datensatzes ENIG.alt. Darin sind alle bereinigten Daten enthalten, die mit dem Nickelbad NIC und dem Goldbad ATS korrespondieren.

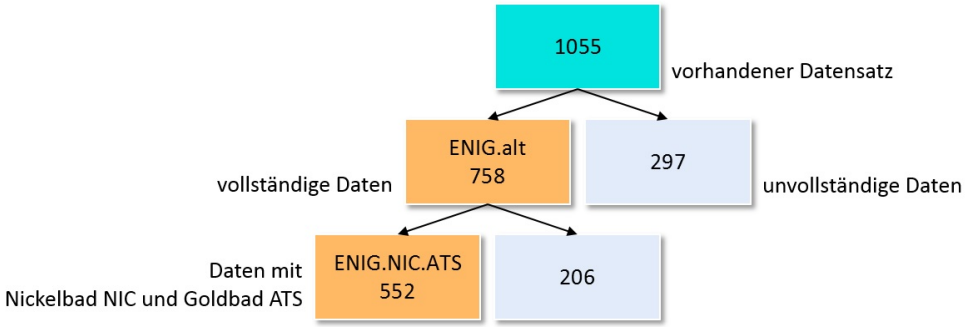

Abbildung 4.3: Historische Datenbasis. Zwei Datensätze (ENIG.alt und ENIG.NIC.ATS) wurden für statistische Analysen herangezogen. Diese sind im Dateiformat CSV (Comma-separated values) abgespeichert.

Mit 297 von 1055 Beobachtungen sind rund 28% der Daten nicht vollständig. Da für eine multiple Regressionsanalyse alle Daten vollständig sein müssen, wurden unvollständige Daten ausgemustert. Eine Aufstellung der fehlenden Werte aller Variablen ist in Tabelle 4.3 enthalten. Abgesehen von der Variable *Schicht* hat die Variable *AuMTO*, der MTO im Goldbad, die meisten fehlenden Werte. Der Prozessparameter mit den wenigsten Lücken ist die Variable *NiTemp*, die Temperatur im Nickelbad.

Tabelle 4.3: Fehlende Werte im vorhandenen Datensatz ($n = 1055$).
Die Tabelle gibt Aufschluss über die Anzahl der Missings jeder Variable.

Nummer	Datum	Schicht					
-	6	363					
NipH	NiTemp	NiKonz	NiMTO	AupH	AuTemp	AuKonz	AuMTO
99	69	101	110	107	76	117	161
Ni1	Ni2	Ni3	Ni4	Ni5	Ni6	Ni7	Ni8
105	104	109	106	106	106	106	106
Au1	Au2	Au3	Au4	Au5	Au6	Au7	Au8
108	108	109	109	109	109	109	109

Bemerkung

- Die Variable Schicht wurde bezüglich der Vollständigkeit nicht berücksichtigt. Da sie in die Analysen nicht miteinbezogen wurde, hätte dies eine unnötige Datenreduktion zur Folge. (In 363 Fällen fehlt die Angabe, siehe Tabelle 4.3).

- Eine Analyse der Missings wurde durchgeführt. Im Hinblick auf die Goldschichtstärke wurden keine Werten systematisch ausgespart. Unvollständige Beobachtungen beinhalten sowohl kleine als auch große Goldschichtstärken.

Messfehler

Einen weiteren Datenqualitätsverlust stellen jegliche Messungenauigkeiten verschiedener Messgeräte dar. Für zukünftige Analysen sollten die unterschiedlichen Messgeräte auf ihre Genauigkeit hin überprüft und nur ein Messgerät pro Parameter verwendet werden.

4.2 Explorative Analyse der historischen Daten

Die explorative Analyse (siehe Kapitel 3.1, Explorative Datenanalyse) ist ein statistisches Instrument zur Untersuchung vorliegender Daten. Unter Einsatz von geeigneten Darstellungen und Berechnungen werden Daten nach Mustern und Zusammenhängen untersucht. Einzelne Variablen werden mittels Kennwerten beschrieben und grafisch dargestellt. Korrelationsmatrizen geben Aufschluss über den linearen Zusammenhang paarweiser Variablen, und anhand Abbildungen unter Einsatz von Farben, Symbolen und unterschiedliche Größen, können Muster zwischen mehreren Variablen identifiziert werden.

4.2.1 Univariate Analyse der historischen Daten

Abbildung 4.4 zeigt einen Überblick aller gemittelten Goldschichtstärken im Zeitverlauf. Allgemein ist in der ersten Hälfte des Beobachtungszeitraums eine fallende Tendenz der mittleren Goldschichtstärke zu erkennen, wohingegen in der zweiten Hälfte des Beobachtungszeitraums ein leicht positiver Trend zu verzeichnen ist. Der Spezifikationsbereich der Goldschichtstärke ist zwischen 60 nm und 120 nm gegeben. Der Idealbereich erstreckt sich zwischen 60 nm und 80 nm. Die meisten Daten liegen im Spezifikationsbereich, einige wenige Schichtstärken liegen außerhalb.

Betrachtet man die Goldschichtstärken getrennt nach den unterschiedlichen Bad-Typen (vgl. Tabelle 4.2), so kann anhand der Boxplots gesehen werden, dass im Zuge der Bad-Kombination CNN/ATS deutlich größere Schichtstärken produziert wurden, als bei den anderen zwei Kombinationen. Die Goldschichtstärken der Kombination CNN/ATS sind fast alle im Spezifikationsbereich, fallen aber aus dem Idealbereich, der Bereich zwischen 60 nm und 80 nm, heraus. Die mittleren 50% der Daten der Kombinationen NIC/ATS und NIC/AU unterschieden sich nicht wesentlich, jedoch ist die Spannweite der Daten der Kombination NIC/AU kleiner.

Dabei ist zu beachten, dass die Anzahl der Beobachtungen der jeweiligen Kombinationen nicht einheitlich ist. Die meisten Beobachtungen mit rund 73% korrespondieren mit dem Nickelbad NIC und dem Goldbad ATS. Rund 16% der Daten fallen auf die

Kombination CNN/ATS und ca. 11% der Daten wurden während der Kombination NIC/AU beobachtet.

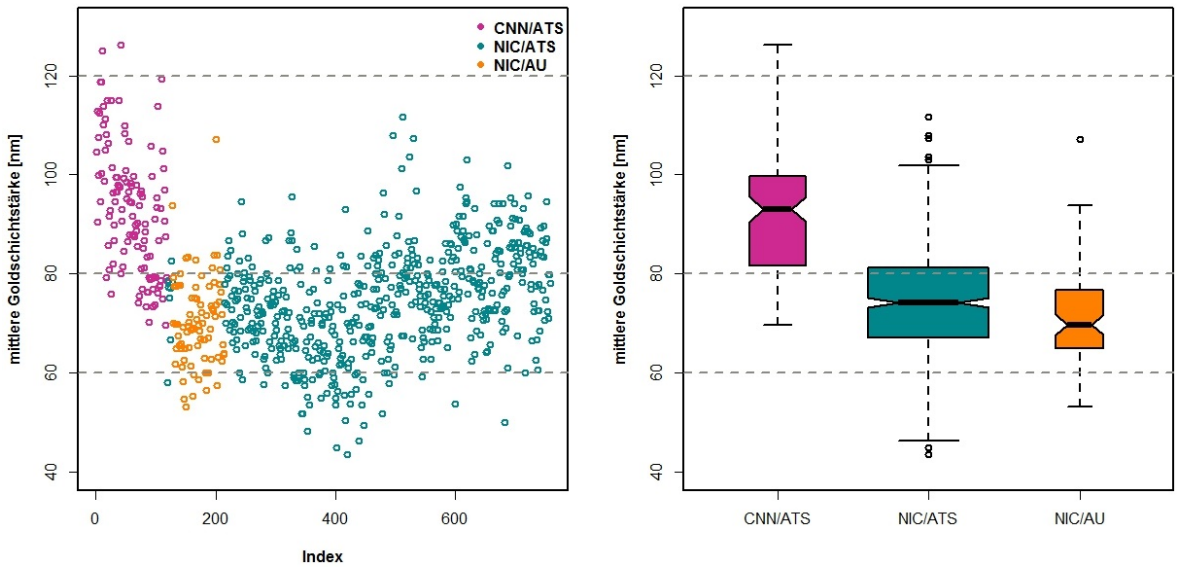

Abbildung 4.4: Mittlere Goldschichtstärke im Zeitverlauf und Boxplotserie der mittleren Goldschichtstärke getrennt nach Bad-Typ (ENIG.alt). Unterschiede in den Schichtstärken bezüglich der Verwendung verschiedener Bad-Typen sind erkennbar.

Ebenso kann die Nickelschichtstärke über die Zeit betrachtet werden (Abbildung 4.5).

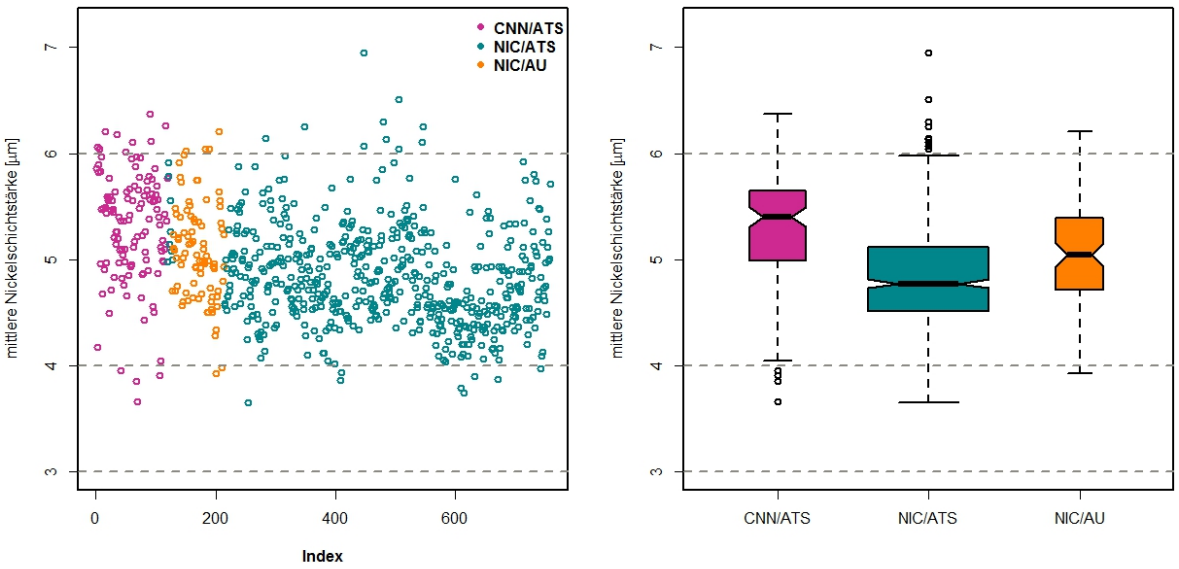

Abbildung 4.5: Mittlere Nickelschichtstärke im Zeitverlauf und Boxplotserie der mittleren Nickelschichtstärke getrennt nach Bad-Typ (ENIG.alt). Die Schichtstärken sind bezüglich der Verwendung verschiedener Bad-Typen unterschiedlich verteilt.

Hierbei lässt sich, im Gegensatz zur mittleren Goldschichtstärke, nur ein schwacher Trend ausmachen. Fast alle Daten befinden sich im Spezifikationsbereich, der zwischen 3 μm und 6 μm liegt. Einige Werte liegen oberhalb des Spezifikationsbereichs, nach unten hingegen sind keine Ausreißer zu verzeichnen. Verkleinert man den Spezifikationsbereich auf den Idealbereich, der zwischen 4 μm und 6 μm ist, so liegen nur wenige Werte unterhalb der Grenze.

Der Unterschied der Nickelschichtstärken getrennt nach Bad-Typen ist gering. Alle drei Kombinationen haben eine ähnliche Spannweite, nur in den mittleren 50% der Daten ergibt sich ein geringer Unterschied. Hier sind die mittleren 50% der Nickelschichtstärken in der Kombination NIC/ATS am kleinsten.

Eine weitere Methode der statistischen Analyse stellen Kennzahlen dar. Die wichtigsten Kennzahlen aller Prozessparameter und der mittleren Schichtstärken über den gesamten Beobachtungszeitraum sind in Tabelle 4.4 aufgelistet.

Bemerkung

- Unterschiedliche Bad-Typen ergeben unterschiedliche Kennzahlen. Teilt man die Daten anhand der Bad-Typen in Untergruppen, so variieren die Kennzahlen stark.

Tabelle 4.4: Deskriptive Statistik der Prozessparameter und mittleren Schichtstärken im historischen Datensatz (ENIG.alt). Die Kennzahlen sind: Minimum, 1. Quantil, Median, Mittelwert, 3. Quantil, Maximum, Standardabweichung.

	Min	$q_{0.25}$	$q_{0.5}$	MW	$q_{0.75}$	Max	Stdabw
NipH	4.00	4.60	4.80	4.74	4.90	5.30	0.18
NiTemp	78.00	82.00	82.00	82.77	82.00	87.00	1.94
NiKonz	4.62	4.95	5.02	5.17	5.11	6.64	0.40
NiMTO	0.00	1.00	1.90	2.03	3.00	6.40	1.32
AupH	4.90	5.40	5.70	5.62	5.80	6.50	0.37
AuTemp	74.00	82.00	84.00	83.97	87.00	90.00	3.15
AuKonz	0.52	0.64	0.69	0.72	0.74	1.18	0.12
AuMTO	0.00	3.00	6.25	6.32	9.60	13.80	3.89
MWNi	3.65	4.56	4.91	4.94	5.26	6.95	0.51
MWAu	43.66	68.41	76.07	76.93	83.88	126.25	12.85

Bemerkung

- Die Spannweiten einzelner Variablen sind ein Produkt langjähriger Erfahrung im Einstellungsprozess. Da die Datenerhebung während der Produktion erfolgte, wurden nur spezielle Parametereinstellungen verwendet.

4.2.2 Multivariate Analyse der historischen Daten

Als nächster Schritt in der explorativen Analyse gilt es den Zusammenhang zwischen den Variablen zu erforschen. Aufschluss über den linearen Zusammenhang jeweils

zweier Variablen gibt die Korrelationsmatrix. Diese wurde mit der **R**-Funktion *corrplot.mixed* aus dem Paket *corrplot* gezeichnet (vgl. Wei, 2015) und ist in Abbildung 4.6 dargestellt.

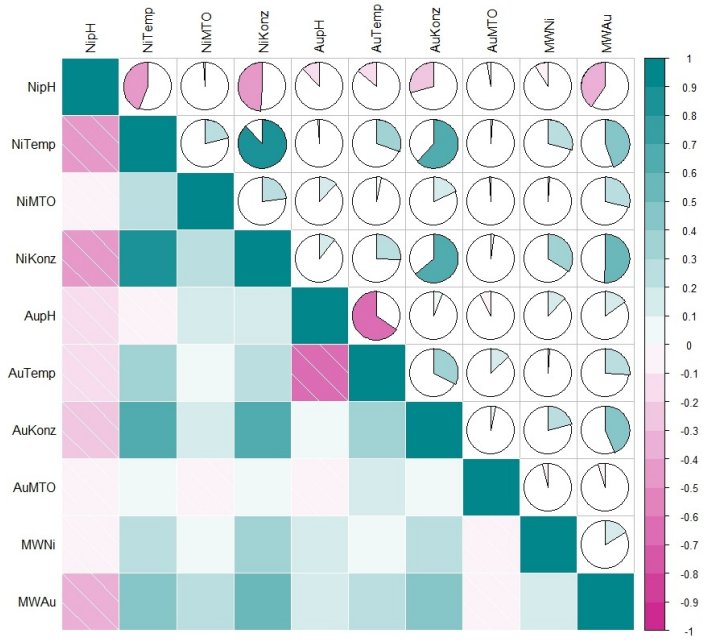

Abbildung 4.6: Korrelationsmatrix der Prozessparameter und mittleren Nickel- und Goldschichtstärken im historischen Datensatz (ENIG.alt). Türkis weist auf eine hohe positive Korrelation hin, weiß auf Unkorreliertheit und pink auf eine hohe negative Korrelation.

Dabei stellt das linke untere Dreieck der Abbildung die Korrelation farbig dar. Je dunkler die Farbe, desto stärker ist die Korrelation. Eine positive Korrelation wird durch einen türkisen Farbton und eine negative Korrelation durch einen pinken Farbton symbolisiert. Im rechten oberen Dreieck der Grafik lässt sich der Wert der Korrelation ablesen. Bei einer positiven Korrelation (türkis) wird die Uhr im Uhrzeigersinn gelesen, wobei eine volle Uhr einen Wert von 1 darstellt. Im Gegensatz dazu wird die Uhr bei einer negativen Korrelation (pink) gegen den Uhrzeigersinn gelesen.

Den stärksten Zusammenhang mit der mittleren Goldschichtstärke haben die Variablen *NiKonz*, *NiTemp*, *AuKonz* und *NipH*. Den schwächsten Zusammenhang mit der mittleren Goldschichtstärke hat hingegen die Variable *AuMTO*.

Generell betrachtet, sind die Zusammenhänge zwischen den Variablen gering. Auffallend ist nur der positive Zusammenhang zwischen der Konzentration und der Temperatur im Nickelbad, der negative Zusammenhang zwischen dem pH-Wert und der Temperatur im Goldbad und die positiven Zusammenhänge zwischen den Konzentrationen beider Bäder, beziehungsweise zwischen der Temperatur im Nickelbad und der Konzentration im Goldbad. Jedoch ändern sich die Zusammenhänge der Variablen, wenn die Daten getrennt nach Bad-Typ analysiert werden.

Abbildung 4.7 stellt die Korrelationsmatrizen der drei verschiedenen Teildatensätze dar, wenn die Daten anhand der verschiedenen Kombinationen von Bad-Typen aufgeteilt werden.

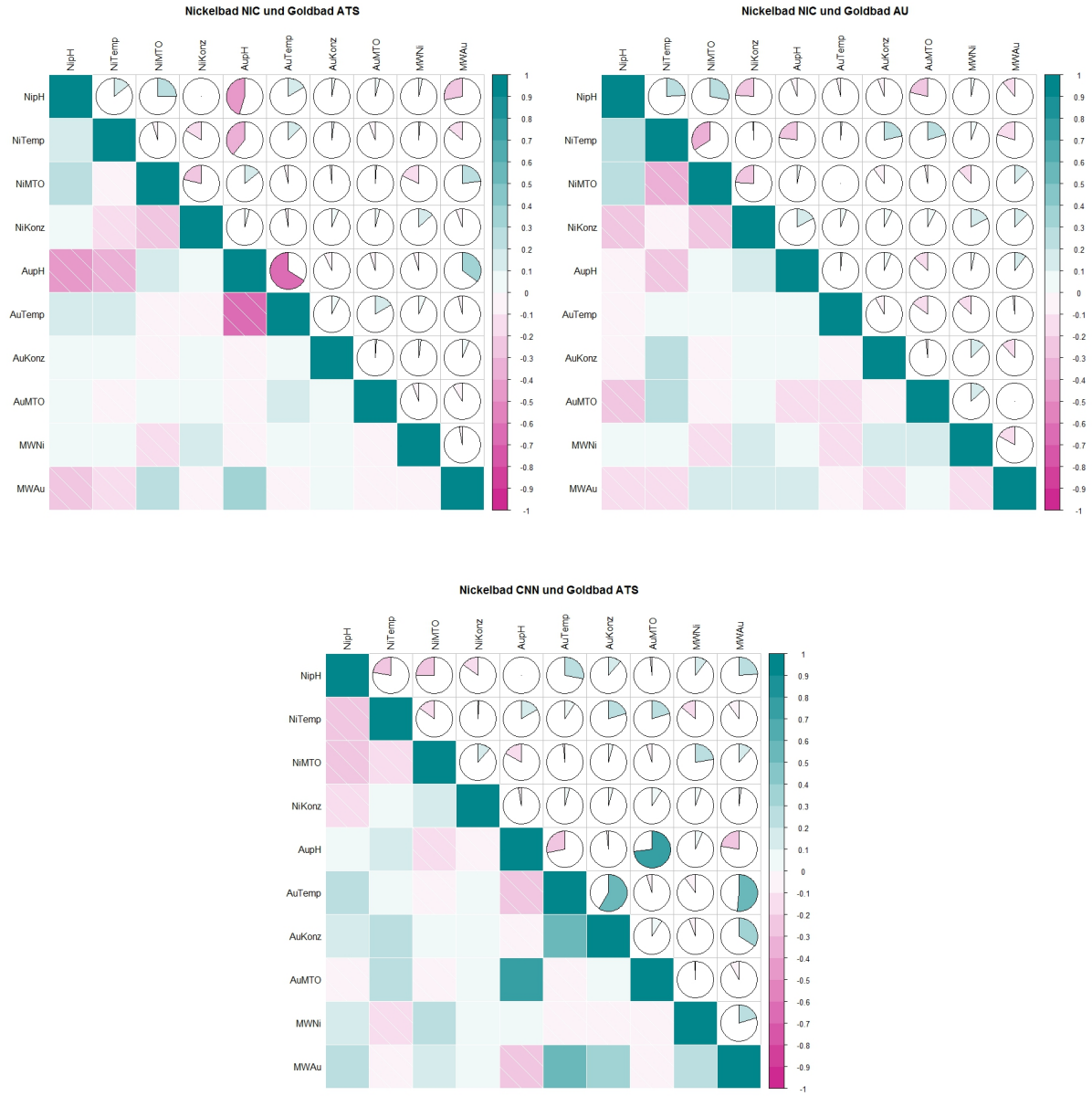

Abbildung 4.7: Korrelationsmatrix der Prozessparameter und mittleren Nickel- und Goldschichtstärken in der verschiedenen Teildatensätzen des historischen Datensatzes. Die drei Teildatensätze entstehen durch Aufspaltung der Daten in die unterschiedlichen Kombinationen von Bad-Typen.

Werden alle Daten zusammen begutachtet, ergibt sich zum Beispiel ein starker positiver Zusammenhang zwischen der Konzentration und der Temperatur im Nickelbad. Werden die Daten getrennt nach Nickelbad-Typ und Goldbad-Typ betrachtet, sind

die Zusammenhänge in allen drei Teildatensätzen kaum vorhanden. Eine andere Situation ergibt sich bei Betrachtung des pH-Wertes und der Temperatur im Goldbad. Die Korrelation der beiden Variablen im Teildatensatz der Kombination Nickelbad NIC und Goldbad ATS ist ähnlich groß, wie über alle Daten. In der Kombination Nickelbad CNN und Goldbad ATS ist der Zusammenhang deutlich geringer. Hingegen ist im Teildatensatz, der mit dem Nickelbad NIC und Goldbad AU korrespondiert, kein Zusammenhang zu erkennen. Demnach ist die Information der unterschiedlichen Bad-Typen wesentlich und muss in die weiteren Analysen miteinbezogen werden.

Bemerkung

- Die Beurteilung von Korrelationskoeffizienten ohne Betrachtung der dahinterstehenden Daten ist nicht sinnvoll. Ebenso können fehlende Informationen zu falschen Interpretationen führen.

4.3 Regressionsanalyse der historischen Daten

Mit Hilfe der linearen Regressionsanalyse soll der Zusammenhang zwischen der mittleren Goldschichtstärke und den Prozessparametern modelliert werden. Dazu wurden einerseits der gesamte historische Datensatz ENIG.alt als auch der Teildatensatz ENIG.NIC.ATS herangezogen.

4.3.1 Modell für die mittlere Goldschichtstärke

Basierend auf der explorativen Analyse wurde ein Regressionsmodell (siehe Kapitel 3.2, Lineare Regression) für die mittlere Goldschichtstärke aufgestellt. Das Ziel dabei ist, das Modell mit möglichst wenigen der verfügbaren Parametern ausreichend gut zu modellieren, vgl. Faraway, 2004. Um die beste Teilmenge an Prädiktoren für das Modell zu finden, wurden statistische Modellkenngrößen herangezogen, wie Akaike´s Informationskriterium (AIC), das Bayes´sche Informationskriterium (BIC), Mallows-C_p-Statistik und das adjustierte Bestimmtheitsmaß (R^2_{adj}). Eine Beschreibung der Kenngrößen findet sich in Kapitel 3.6, Modellbildung.

Die Auswahl der Variablen für das Regressionsmodell basiert auf der schrittweisen Regression und auf einer vollständigen Modellselektion anhand der Kennzahlen BIC, C_p und R^2_{adj}. Mit Hilfe der **R**-Funktion *regsubsets* des Pakets *leaps* und der **R**-Funktion *step* wurde die Variablenselektion durchgeführt (vgl. Lumley, 2015). Dabei sucht die Funktion *regsubsets* nach der besten Teilgruppe an Prädiktoren zur Beschreibung der Zielvariablen, wobei das Kriterium zur Bewertung der Prädiktoren-Teilgruppen ausgewählt werden kann, während die Funktion *step* eine schrittweise Regression durchführt. Für genaue Informationen der Funktionen *regsubsets* und *step* wird auf die Ausführungen in Lumley, 2015 und das Hilfe-Handbuch von **R** verwiesen.

Abbildung 4.8 zeigt die Veränderungen des BIC für unterschiedliche Gruppen von Prädiktoren. Eine Zeile entspricht einem Modell, wobei die nicht-weißen Markierungen mit dem BIC korrespondieren und die zugehörigen Variablen als Prädiktoren

im Modell enthalten sind. Dabei ist bei den untersten vier Zeilen eine deutliche Modellverbesserung erkennbar. Dies zeigt auch Abbildung 4.9. Die restlichen Modelle unterscheiden sich bezüglich des zugehörigen Informationskriteriums sehr wenig.

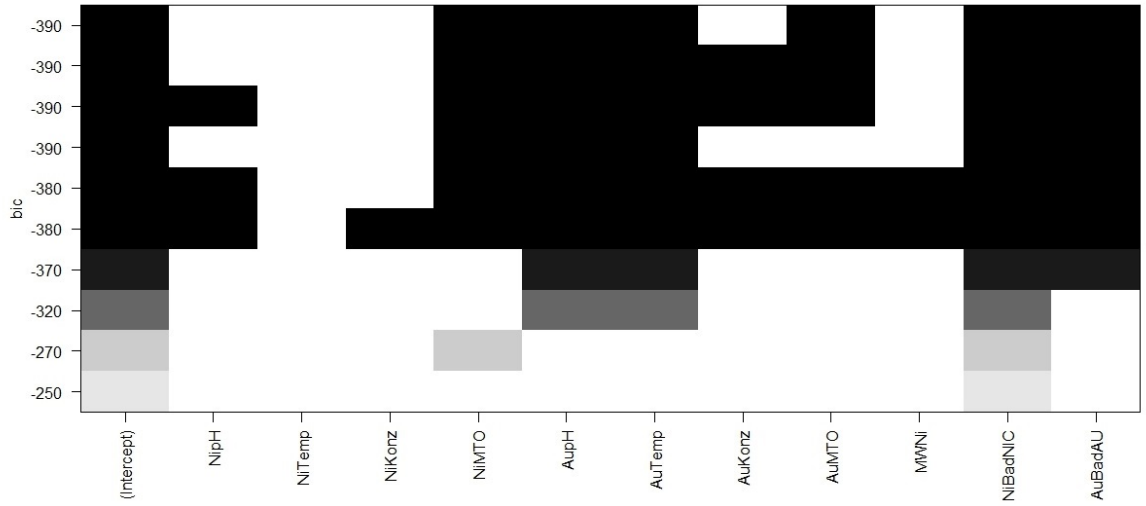

Abbildung 4.8: Variablenselektion anhand des Bayes´schen Informationskriteriums im historischen Datensatz (ENIG.alt). Jede Zeile entspricht einem Modell, in dem die Variablen mit nicht-weißen Markierungen als Prädiktoren enthalten sind.

Aufgrund der Analyse des BIC, des C_p-Wertes und des adjustierten R^2, wurde ein Regressionsmodell erstellt. Das vermeintlich beste Modell ist das Modell mit dem kleinsten BIC, einem kleinen C_p-Wert, der gleich p ergibt, wobei p die Anzahl der Prädiktoren inklusive Intercept ist, und das adjustierte Bestimmtheitsmaß möglichst groß ist. Die Entwicklung dieser Kriterien zu den Modellen aus Abbildung 4.8 stellt Abbildung 4.9 dar.

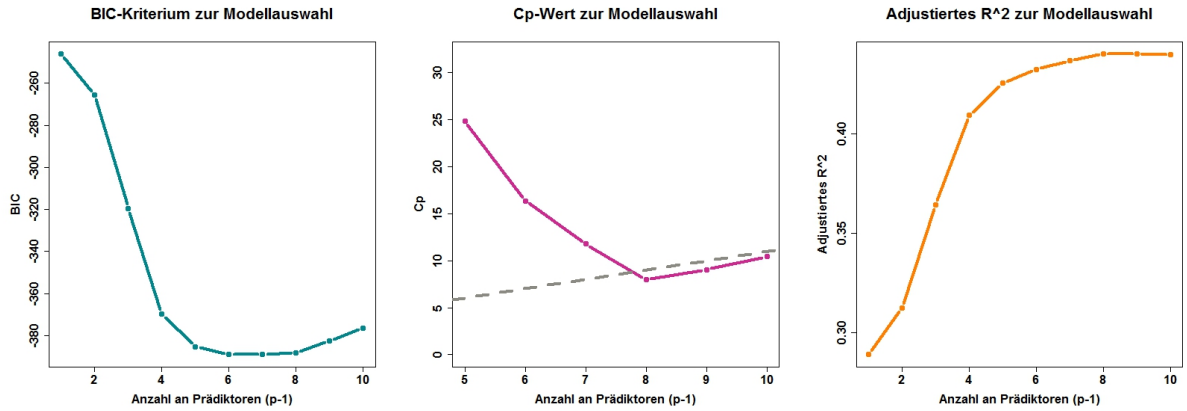

Abbildung 4.9: Informationskriterien zur Modellauswahl im historischen Datensatz (ENIG.alt). Für das beste Modell ist das BIC minimal, die C_p-Statistik gleich p und R^2_{adj} möglichst groß.

Die Optimierung aller drei Kriterien liefert dabei leicht unterschiedliche Ergebnisse. Anhand des BIC sollen sechs Prädiktoren für das beste Modell verwendet werden, während laut C_p-Statistik und R^2_{adj} das Modell mit acht Prädiktoren das bestmögliche ist. Da das Regressionsmodell die Beziehung der Variablen mit möglichst wenigen Prädiktoren möglichst gut beschreiben soll, wurde das Modell mit sechs Parametern gewählt.

Eine Zusammenfassung des Regressionsmodells liefert Tabelle 4.5. Die Anzahl der Prädiktoren hat sich demnach von zehn möglichen Prädiktoren (acht Prozessparameter, Nickelbad-Typ und Goldbad-Typ) auf sechs reduziert.

Tabelle 4.5: Lineares Regressionsmodell im historischen Datensatz (ENIG.alt). Im Modell sind nur Haupteffekte enthalten.

$MWAu^{1/3} \sim$ NiBad+AuTemp+AupH+AuBad+NiMTO+AuMTO					
	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	0.184729	0.384933	0.480	0.63144	
NiBadNIC	-0.177803	0.021102	-8.426	< 2e-16	***
AuTemp	0.026213	0.003065	8.554	< 2e-16	***
AupH	0.358066	0.029930	11.963	< 2e-16	***
AuBadAU	-0.240900	0.028834	-8.355	3.16e-16	***
NiMTO	0.024274	0.005125	4.736	2.61e-06	***
AuMTO	-0.004935	0.001683	-2.932	0.00347	**
Observations	758				
R^2	0.423				
Adjusted R^2	0.419				
Residual Std. Error	0.178 (df = 751)				
F Statistic	91.934 *** (df = 6; 751)				
Signif. codes:	*** p<0.001 ** p<0.01 * p<0.05 . p<0.1				

Neben der Temperatur und dem pH-Wert im Goldbad und der MTOs beider Bäder, spielen die Bad-Typen eine wichtige Rolle, was anhand der explorativen Analyse schon zu erkennen war. Das adjustierte R^2 liegt bei 0.419 und der Residual Standard Error bei 0.178.

Die Response wurde auf Basis eines Boxcox-Tests transformiert. Durch die Transformation werden die Modellannahmen besser erfüllt (siehe Transformationen, 3.5).

Im Anschluss an die Modellierung eines Regressionsmodells ist es unerlässlich, die Voraussetzungen des Regressionsmodells zu überprüfen. Der Überprüfung des Modells und der vorausgesetzten Annahmen dienen die Grafiken in Abbildung 4.10. So sollten die Residuen keinen systematischen Trend aufweisen, die standardisierten Residuen normalverteilt sein, Varianzhomogenität der Residuen herrschen und die Regressionskoeffizienten nicht von einzelnen Beobachtungen abhängen (siehe Kapitel Modelldiagnose, 3.4).

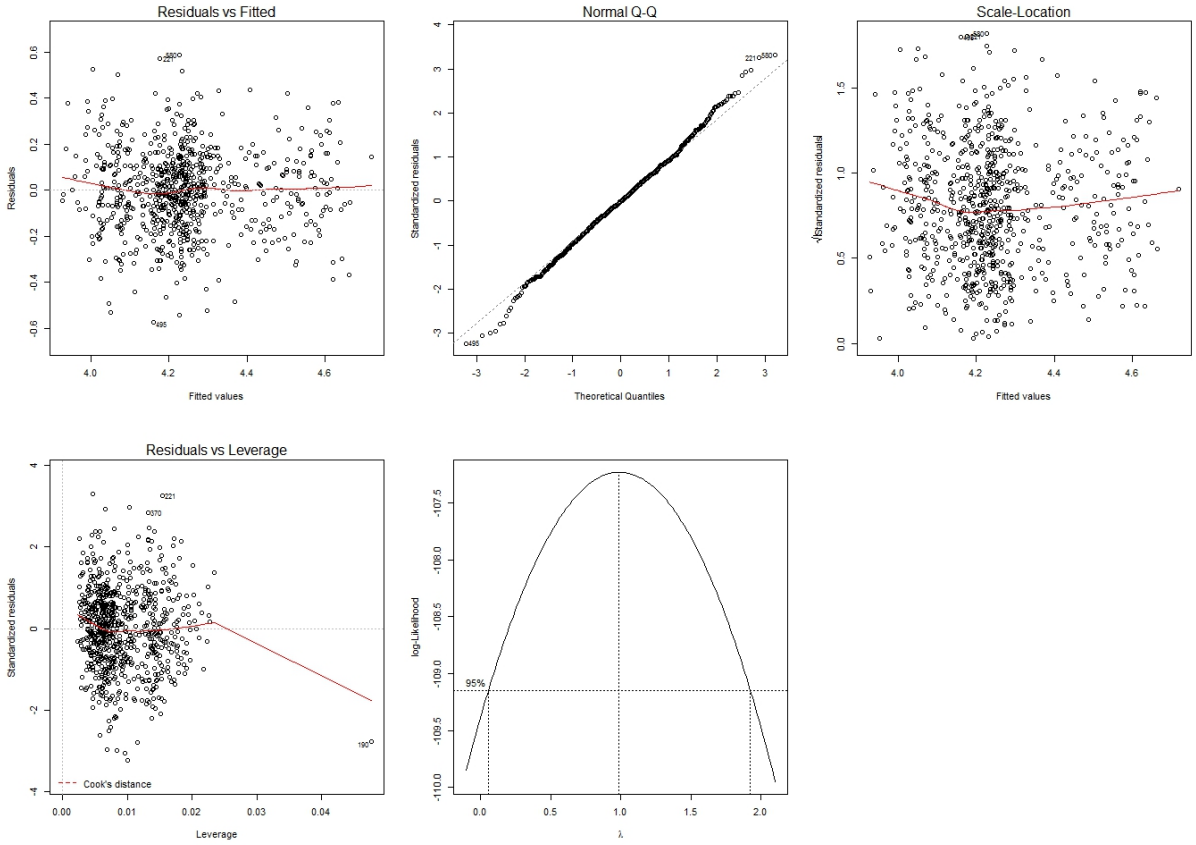

Abbildung 4.10: Modelldiagnoseplots im historischen Datensatz (ENIG.alt). Unter Zuhilfenahme grafischer Methoden können die Modellannahmen überprüft werden.

Die erste Grafik zeigt einen Scatterplot der Residuen gegen die vorhergesagten Werte. Dabei ist kein systematischer Trend identifizierbar. Die nächste Grafik zeigt einen Q-Q-Plot, der eine Normalverteilung der standardisierten Residuen erkennen lässt, da die Punkte gut auf der Referenzlinie liegen. In der dritten Grafik sind die vorhergesagten Werte gegen die Wurzeln aus dem Betrag der standardisierten Residuen aufgetragen. Die zufällige Verteilung der Datenpunkte weist auf eine Varianzhomogenität der Residuen hin. Ob Beobachtungen weit weg vom Zentrum der Daten und/oder besonders einflussreich sind, kann mit der vierten Grafik beantwortet werden. Dabei sticht besonders die 190. Beobachtung heraus, die aber nicht einflussreich ist. Die letzte Grafik überprüft, ob eine Transformation der Response notwendig ist. Da diese bereits transformiert wurde, ist keine weitere Transformation angezeigt. Ein Indiz dafür ist, dass im 95% Konfidenzintervall die Eins als mögliche Hochzahl für eine Transformation enthalten ist.

4.3.2 Modell für die mittlere Goldschichtstärke im Teildatensatz

Wie im vorigen Regressionsmodell herausgefunden wurde, spielen sowohl der Nickelbad-Typ als auch der Goldbad-Typ eine signifikante Rolle. Die Vorhersagen sind

demnach je nach Bad-Typ unterschiedlich. Um ein Modell zu erstellen, welches mit den Modellen des neuen Datensatzes vergleichbar ist, wurde ein Teildatensatz mit der Bad-Kombination NIC/ATS betrachtet. Dieser wurde mit ENIG.NIC.ATS gekennzeichnet und beinhaltet 552 Beobachtungen, die mit dem Nickelbad NIC und dem Goldbad ATS korrespondieren, vgl. Abbildung 4.3.

Um die linearen Zusammenhänge zu erforschen, wurde eine Korrelationsmatrix erzeugt. Abbildung 4.11 stellt die Korrelationsmatrix der Variablen im Teildatensatz dar. Ein Vergleich der Korrelationsmatrizen des gesamten Datensatzes (Abbildung 4.6) und des Teildatensatzes (Abbildung 4.11) zeigt einige Unterschiede.

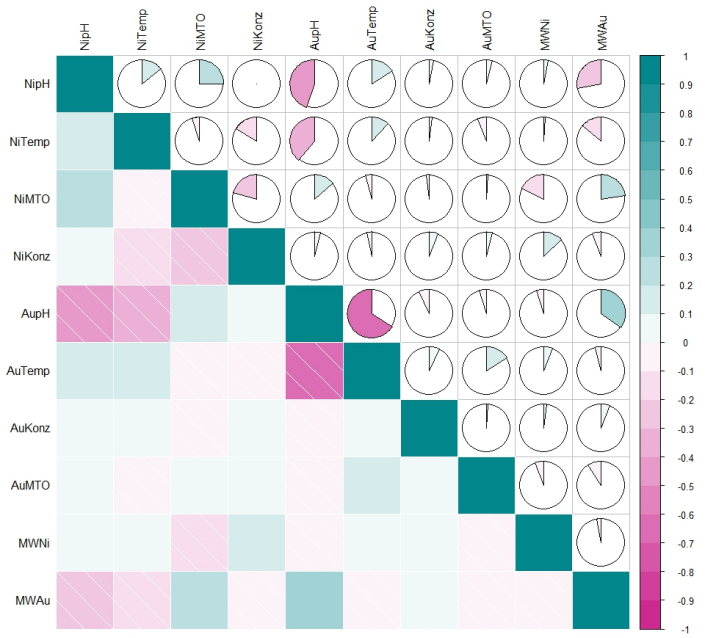

Abbildung 4.11: Korrelationsmatrix der Variablen im reduzierten historischen Datensatz (ENIG.NIC.ATS). Türkis weist auf eine hohe positive Korrelation hin, weiß auf keine Korrelation und pink auf eine hohe negative Korrelation.

So ist zum Beispiel die Temperatur und die Konzentration im Nickelbad im gesamten Datensatz ENIG.alt stark positiv korreliert und im Teildatensatz ENIG.NIC.ATS schwach negativ korreliert.

Für das Regressionsmodell wurden zusätzlich zu den Haupteffekten alle möglichen Wechselwirkungen zwischen zwei Variablen betrachtet. Die Funktion *regsubsets* liefert als Basis für die Variablenselektion Abbildung 4.12 und Abbildung 4.13. In Abbildung 4.12 ist die Entwicklung des BIC für die unterschiedlichen Gruppen an Prädiktoren dargestellt. Eine nicht-weiße Markierung symbolisiert die Zugehörigkeit zur Gruppe der Prädiktoren. Dabei fällt in Abbildung 4.12 auf, dass Wechselwirkungen oft ohne die zugehörigen Haupteffekte im Modell enthalten sind. Diese sollten jedoch nur gemeinsam mit den zugehörigen Haupteffekten in ein Modell inkludiert werden.

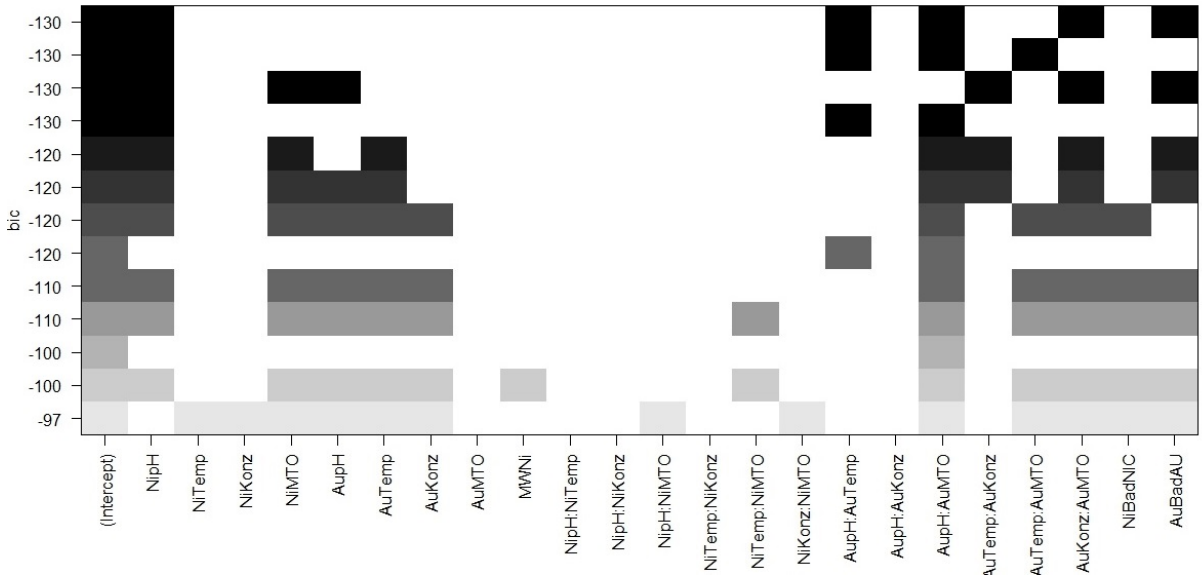

Abbildung 4.12: Variablenselektion anhand des Bayes´schen Informationskriteriums im reduzierten historischen Datensatz (ENIG.NIC.ATS). Jede Zeile entspricht einem Modell, in dem die Variablen mit nicht-weißen Markierungen als Prädiktoren enthalten sind.

Bemerkung

- Terme höherer Ordnung, wie zB. Interaktionen, sollten nur dann in einem Regressionsmodell enthalten sein, wenn die zugehörigen Variablen als Haupteffekt im Modell sind, vgl. Faraway, 2004.

Die zugehörigen Grafiken, die die Veränderungen des BIC, des C_p-Wertes und des adjustierten R^2 zeigen, liefert Abbildung 4.13.

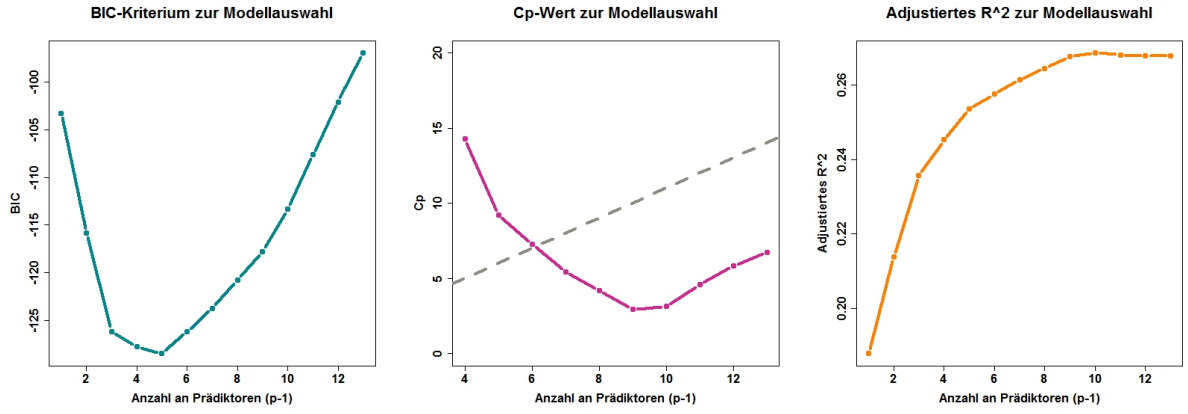

Abbildung 4.13: Informationskriterien zur Modellauswahl im reduzierten historischen Datensatz (ENIG.NIC.ATS). Für das beste Modell ist das BIC minimal, die C_p-Statistik gleich p und R^2_{adj} möglichst groß.

Dabei ergibt sich für das optimale Modell je nach Kriterium eine Anzahl von fünf, sechs bzw. zehn Prädiktoren.

Falls eine Wechselwirkung in ein Modell aufgenommen wird, sollten auf jeden Fall alle zugehörigen Haupteffekte im Modell berücksichtigt werden. So wurden nach der Modellauswahl, basierend auf obigen Informationskriterien, die Haupteffekte zu signifikanten Wechselwirkungen hinzugefügt. Mittels stufenweiser Regression wurde das erweiterte Modell wiederum reduziert, indem alle nicht signifikanten Prädiktoren entfernt wurden. Dabei kristallisierte sich folgendes Modell heraus (Tabelle 4.6):

Tabelle 4.6: Lineares Regressionsmodell im reduzierten historischen Datensatz (ENIG.NIC.ATS). Neben den Haupteffekten sind zwei Wechselwirkungen im Modell enthalten.

MWAu∼
NiMTO+AuTemp+NipH+AupH+AuKonz+AupH:AuTemp+AuTemp:AuKonz

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	1527.6662	374.1607	4.083	5.12e-05	***
NiMTO	2.0789	0.3670	5.665	2.39e-08	***
AuTemp	-17.5599	4.4303	-3.964	8.37e-05	***
NipH	-12.2099	3.2270	-3.784	0.000172	***
AupH	-188.6858	57.8277	-3.263	0.001172	**
AuKonz	-654.9104	246.8922	-2.653	0.008220	**
AuTemp:AupH	2.3914	0.6816	3.509	0.000488	***
AuTemp:AuKonz	7.9692	2.9373	2.713	0.006876	**
Observations	552				
R^2	0.270				
Adjusted R^2	0.261				
Residual Std. Error	9.086 (df = 544)				
F Statistic	28.783 *** (df = 7; 544)				
Signif. codes:	*** p<0.001 ** p<0.01 * p<0.05 . p<0.1				

Die signifikanten Parameter haben sich im Vergleich zum vorigen Modell, Tabelle 4.5, verändert. So ist nun zusätzlich die Konzentration im Goldbad und der pH-Wert im Nickelbad im Modell enthalten, wie auch die Wechselwirkung zwischen der Temperatur und dem pH-Wert, sowie der Temperatur und der Konzentration im Goldbad. Andererseits ist der MTO im Goldbad nicht mehr signifikant. Das adjustierte R^2 ist von 0.419 auf 0.261 gesunken. So wird durch das Modell nicht mehr so viel erklärt wie bei Verwendung des gesamten Datensatzes.

Bemerkung

- Da die Responsevariablen beider Regressionsmodelle nicht identisch transformiert sind, sind die geschätzten Koeffizienten und die standardisierten Residuals Errors der Modelle der Datensätze ENIG.alt und ENIG.NIC.ATS nicht direkt vergleichbar, vgl. Tabelle 4.5 und 4.6.

Die Analyse des Modells mittels Diagnoseplots ist in Abbildung 4.14 abgebildet.

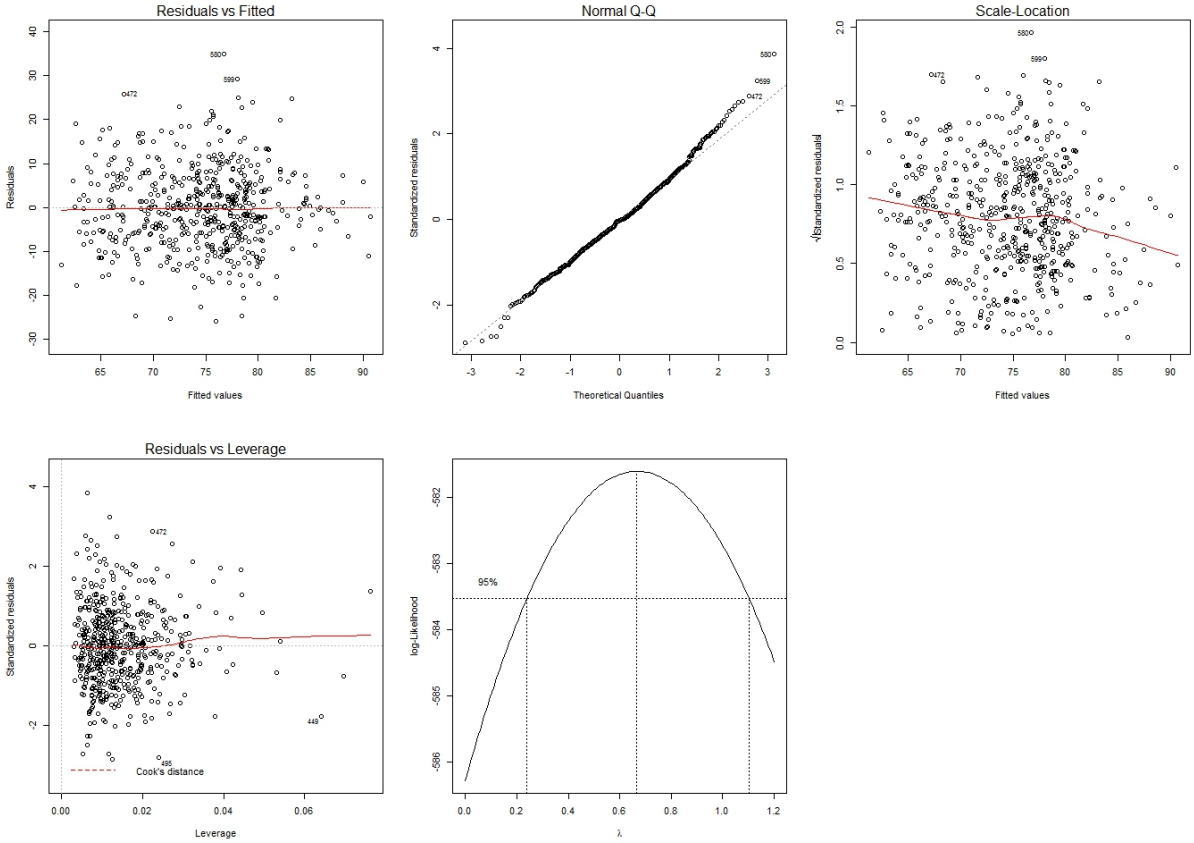

Abbildung 4.14: Modelldiagnoseplots im reduzierten historischen Datensatz (ENIG.NIC.ATS). Unter Zuhilfenahme grafischer Methoden werden die Modellannahmen überprüft. Dabei scheinen keine Annahmen verletzt zu sein.

Auch hier sind keine Auffälligkeiten sichtbar. Im Scatterplot der Residuen gegen die vorhergesagten Werte sind keinerlei Strukturen wahrnehmbar. Der Q-Q-Plot der standardisierten Residuen widerspricht nicht der Annahme der Normalverteilung. Die Annahme der Varianzhomogenität der Residuen wird mittels Scale-Location-Plot nicht verworfen und keine Beobachtungen beeinflussen das Modell übermäßig stark. Die letzte Grafik weist darauf hin, dass keine weitere Transformation der Response notwendig zu sein scheint.

Zur Visualisierung des Regressionsmodells wurde ein Regressionsbaum gezeichnet. Dieser wird in Abbildung 4.15 dargestellt. Der abgebildete Regressionsbaum wurde mit der **R**-Funktion *ctree* aus dem Paket *party* erstellt. Für Informationen über das Paket wird auf die Ausführungen in Hothorn u. a., 2015, verwiesen.

Ein Baum besteht aus einer Wurzel, Ästen und Blättern. Seine Struktur gibt stufenweise Entscheidungen wieder. Der wichtigste Prozessparameter bildet die Wurzel des Baumes. Sie ist an der Spitze der Grafik zu erkennen und ist in diesem Fall der pH-Wert im Goldbad. Die inneren Knoten stellen die weiteren Variablen dar, nach denen die Daten

aufgespalten werden. Ausgehend von der Wurzel spaltet sich der Baum in zwei Teile auf. Der linke Teil führt direkt zu einem Blatt, während der rechte Teil weitere zwei Knoten besitzt, den MTO und den pH-Wert im Nickelbad. Ausgehend vom MTO im Nickelbad mündet der linke Ast in einem Blatt und der rechte Ast im Knoten pH-Wert im Nickelbad. Die Blätter stellen die zugehörigen Daten anhand eines Boxplots dar.

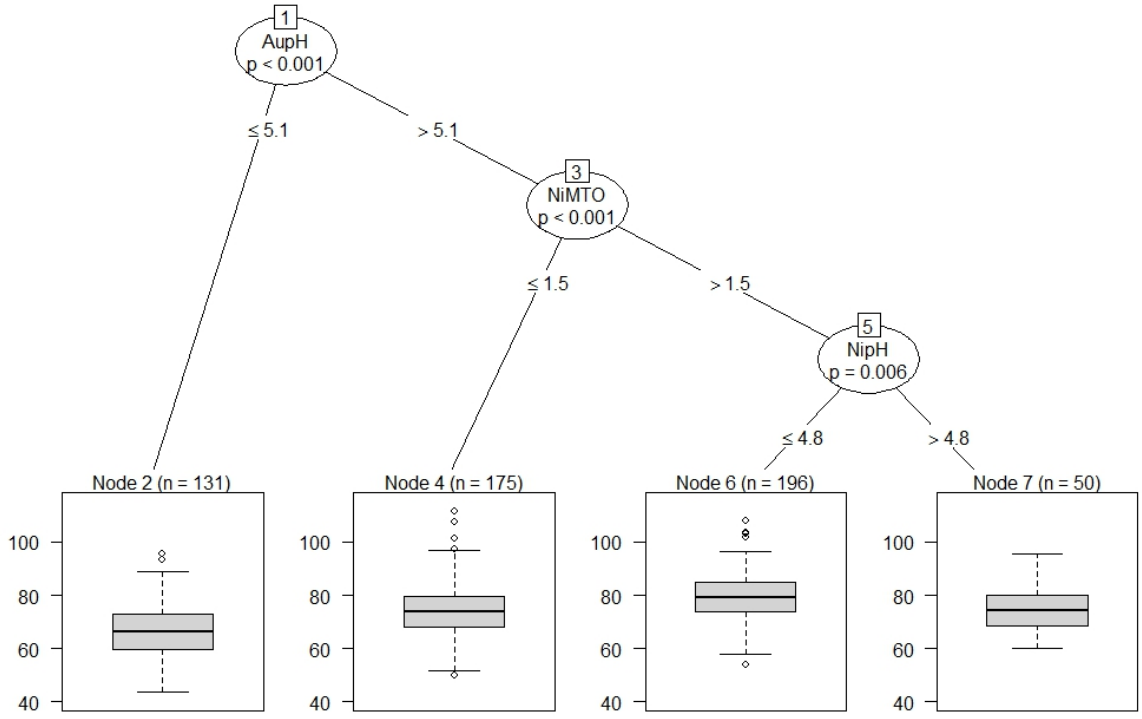

Abbildung 4.15: Regressionsbaum im reduzierten historischen Datensatz (ENIG.NIC.ATS). Die Beobachtungen werden auf Basis der Prädiktoren im Regressionsmodell in Gruppen aufgeteilt. Wird der Baum beginnend bei den Blättern gelesen, ergeben sich je nach Blatt unterschiedliche Parametereinstellungen.

Falls der pH-Wert im Goldbad kleiner oder gleich 5.1 ist, liefert der linke Ast des Regressionsbaums ohne Zwischenknoten ein Blatt. Diese Entscheidung ist im Bezug auf Minimierung der Goldschichtstärke die beste. Alle anderen Entscheidungen führen zu Verteilungen, deren Goldschichtstärken im Mittel größer sind. Trotzdem sind auch bei den übrigen drei Blättern Unterschiede in der Verteilung erkennbar. Sowohl in der Lage der Box als auch der Länge der Tails.

Eine mögliche andere Strategie ist die Vermeidung jener Entscheidungen, die die höchsten Goldschichtstärken zu Folge haben. So eine Entscheidung wäre

(pH-Wert im Goldbad größer als 5.1) und
(MTO im Nickelbad größer als 1.5) und
(pH-Wert im Nickelbad kleiner gleich 4.8).

Diese Parametereinstellungen liefern im Mittel größere Goldschichtstärken als die anderen Entscheidungen.

4.4 Modellierungsdaten

Ausgehend von der Studie des historischen Datensatzes und dessen Qualitätslücken wurde ein neuer Datensatz generiert. Dieser sollte einerseits alle möglichen Prozessparameter beinhalten, die einen möglichen Einfluss auf die Goldschichtstärke haben, und andererseits die geforderte Datenqualität für eine angemessene statistische Auswertung aufweisen.

Schlussendlich wurden die meisten Parameter festgehalten, die im gesamten Prozessverlauf gemessen werden. Zusätzlich wurde der Zeitstempel verfeinert und die Messungen der Nickel- und Goldschichtstärke wurden adaptiert.

4.4.1 Beschreibung des Datensatzes

Die Datenerhebung wurde im Zeitraum vom 26.11.2014 bis 28.02.2015 durchgeführt.

Im Vergleich zum historischen Datensatz wurden in folgenden drei Bereichen Änderungen vorgenommen:

- Intensität der Messungen,
- Anzahl und Art der Prozessparameter und
- Messprozedur der Schichtstärkenmessungen.

Um in relativ kurzer Zeit möglichst viele Daten zu bekommen, wurde die Anzahl der Messungen erhöht. Dazu wurde das Produktionsintervall der Testleiterplatte von 24-stündlich auf 4-stündlich verkürzt. Um eine genaue zeitliche Einordnung der Beobachtungen zu ermöglichen, wurde der Zeitstempel dafür um die Variable Uhrzeit erweitert.

Die Anzahl und Art der Prozessparameter änderte sich folgendermaßen: Alle wesentlichen Prozessparameter, welche die Goldschichtstärke beeinflussen, sollten nach Möglichkeit gemessen werden. Der historische Datensatz wurde demnach um einige Variablen erweitert. Im neuen Datensatz wurden statt der ursprünglichen acht Prozessparameter 34 Parameter gemessen. Dabei wurde das Augenmerk auf das Nickelbad bzw. das Goldbad gelegt. Die bisher gemessenen Variablen wurden bis auf die Soll-Temperatur beibehalten. Die Soll-Temperatur im Nickelbad und im Goldbad wurden durch die Minimaltemperatur und Maximaltemperatur ersetzt. Diese bilden die tatsächliche Temperatur besser ab. Einen Überblick der schlussendlich beobachteten Prozessparameter der Aktivstationen und ihrer Messstellen im gesamten Prozessverlauf gibt Abbildung 4.16.

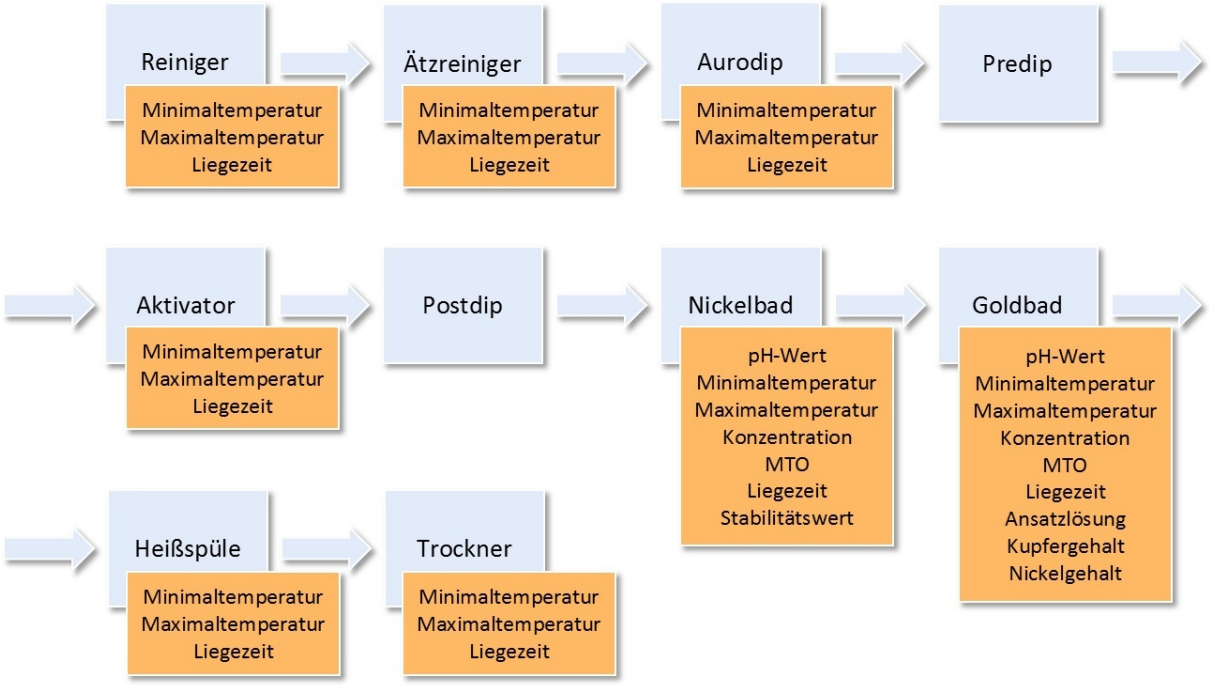

Abbildung 4.16: Variablen im Prozessverlauf im Modellierungsdatensatz (ENIG.neu). Insgesamt wurden 34 Prozessparameter in acht verschiedenen Aktivstationen gemessen.

Die Messprozedur der Schichtstärken wurde im Vergleich zum historischen Datensatz ebenfalls geändert. Statt die Nickel- und Goldschichtstärken achtmal auf beliebigen Pads mittlerer Größe zu messen, wurden nun Messungen auf Pads unterschiedlicher Größe vorgenommen. Wie in Abbildung 4.17 ersichtlich, wurden zwölf Messungen der Schichtstärken durchgeführt. Dafür wurden jeweils vier Messungen auf vordefinierten kleinen, mittleren und großen Pads durchgeführt. Somit konnte auch der Einfluss der Pad-Größe untersucht werden.

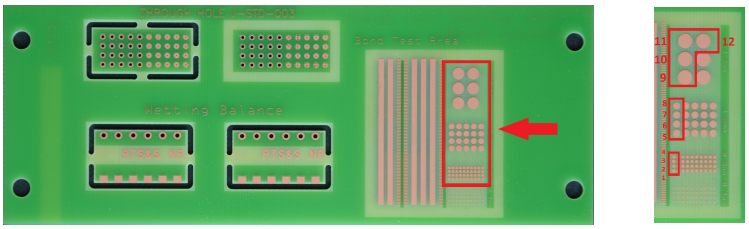

Abbildung 4.17: Testleiterplatte und adaptierte Messprozedur der Schichtstärke. Die Nickel- bzw. Goldschichtstärke wurde auf je vier verschiedenen kleinen/mittleren/großen Pads gemessen. Die Durchmesser der Pads sind 0.8 *mm*, 1.6 *mm* und 3.2 *mm*.

Eine Übersicht aller Parameter des neuen Datensatzes samt ihren Abkürzungen und Einheiten liefert Tabelle 4.7. Für die weiteren Analysen wurden die vier Werte der Schichtstärken für jede Pad-Größe gemittelt.

Tabelle 4.7: Variablen im Modellierungsdatensatz (ENIG.neu). Neben dem Zeitstempel sind pro Beobachtung 40 Parameter, die Nickelschichtstärken und die Goldschichtstärken festgehalten.

	Variable	**Abkürzung**	**Einheit**
Zeitstempel	Nummer	Nummer	
	Datum	Datum	
	Schicht	Schicht	
	Uhrzeit	Uhrzeit	
Reiniger	Minimaltemperatur [4]	Reiniger_Tmin	[°C]
	Maximaltemperatur [5]	Reiniger_Tmax	[°C]
	Liegezeit [6]	Reiniger_Liegezeit	[*min*]
Ätzreiniger	Minimaltemperatur [4]	Microaetzen_Tmin	[°C]
	Maximaltemperatur [5]	Microaetzen_Tmax	[°C]
	Liegezeit [6]	Microaetzen_Liegezeit	[*min*]
Aurodip	Minimaltemperatur [4]	AuroDip_Tmin	[°C]
	Maximaltemperatur [5]	AuroDip_Tmax	[°C]
	Liegezeit [6]	AuroDip_Liegezeit	[*min*]
Aktivator	Minimaltemperatur [4]	AktivierenAu_Tmin	[°C]
	Maximaltemperatur [5]	AktivierenAu_Tmax	[°C]
	Liegezeit [6]	AktivierenAu_Liegezeit	[*min*]
Nickelbad	pH-Wert	NipH	
	Minimaltemperatur [4]	ChemNi_Tmin	[°C]
	Maximaltemperatur [5]	ChemNi_Tmax	[°C]
	Konzentration [7]	NiKonz	[*g*/*l*]
	MTO [8]	NiMTO	
	Liegezeit [6]	ChemNi_Liegezeit	[*min*]
	Stabilitätswert [9]	NiStabi	
Goldbad	pH-Wert	AupH	
	Minimaltemperatur [4]	SudGold_Tmin	[°C]
	Maximaltemperatur [5]	SudGold_Tmax	[°C]
	Konzentration [7]	AuKonz	[*g*/*l*]

Fortsetzung nächste Seite

[4]Die Minimaltemperatur ist die niedrigste Temperatur, die gemessen wurde, während die Testleiterplatte in der Aktivstation war, Breitwieser, 2015.

[5]Die Maximaltemperatur ist die höchste Temperatur, die gemessen wurde, während die Testleiterplatte in der Aktivstation war, Breitwieser, 2015.

[6]Die Liegezeit gibt die Verweildauer der Testleiterplatten in der jeweiligen Aktivstation an, Breitwieser, 2015.

[7]Die Konzentration gibt Auskunft über die Anteile eines bestimmten Stoffes (hier: Nickel bzw. Goldsalz) im gesamten Badvolumen, Breitwieser, 2015.

[8]Der MTO (Metal-Turn-Over) gibt das Alter des Bades an. Er erhöht sich bei jeder Zugabe von Nickel bzw. Goldsalz. Da die Bäder keine unendliche Lebensdauer besitzen, müssen sie regelmäßig erneuert werden. Der MTO wird dabei wieder auf Null gesetzt, Breitwieser, 2015.

[9]Der Stabilitätswert ist eine dimensionslose Zahl, welche Informationen über die Menge des Stabilisators im Nickelbad enthält, Breitwieser, 2015.

Tabelle 4.7 – *Fortsetzung von vorheriger Seite*

	Variable	**Abkürzung**	**Einheit**
	MTO [8]	AuMTO	
	Liegezeit [6]	SudGold_Liegezeit	[min]
	Ansatzlösung [10]	AuAnsatz	[%]
	Kupfergehalt [11]	AuCu	[mg/l]
	Nickelgehalt [12]	AuNi	[mg/l]
Heißspüle	Minimaltemperatur [4]	Heissspuele_Tmin	[$°C$]
	Maximaltemperatur [5]	Heissspuele_Tmax	[$°C$]
	Liegezeit [6]	Heissspuele_Liegezeit	[min]
Trockner	Minimaltemperatur [4]	Trockner_Tmin	[$°C$]
	Maximaltemperatur [5]	Trockner_Tmax	[$°C$]
	Liegezeit [6]	Trockner_Liegezeit	[min]
Sonstiges	Gesamtfläche [13]	Total_Area	[dm^2]
	Kupferfläche [14]	Copper_Area	[dm^2]
Nickelschichtstärken	Pad klein	Ni1K bis Ni4K	[μm]
		MWNiK	[μm]
	Pad mittel	Ni1M bis Ni4M	[μm]
		MWNiM	[μm]
	Pad groß	Ni1G bis Ni4G	[μm]
		MWNiG	[μm]
Goldschichtstärken	Pad klein	Au1K bis Au4K	[μm]
		MWAuK	[nm]
	Pad klein	Au1M bis Au4M	[μm]
		MWAuM	[nm]
	Pad klein	Au1G bis Au4G	[μm]
		MWAuG	[nm]

Bemerkung

- Die dritte Spalte in Tabelle 4.7 gibt die Abkürzungen der Variablen an, die im Weiteren verwendet werden. MWAuK, MWAuM und MWAuG bezeichnen die gemittelten Goldschichtstärken der unterschiedlichen Pad-Größen.

[10] Die Ansatzlösung beinhaltet alle Komponenten, welche für die Stabilität des Goldbads verantwortlich sind. Diese sind vor allem Komplexbildner und organische Stabilisatoren, Breitwieser, 2015.

[11] Bedingt durch die Reaktion, wird immer etwas Kupfer von der Leiterplatte abgelöst. Durch den Kupfergehalt erhält man Informationen über die Belastung des Bades, Breitwieser, 2015.

[12] Bedingt durch die Reaktion, wird immer etwas Nickel von der Leiterplatte abgelöst. Durch den Nickelgehalt erhält man Informationen über die Belastung des Bades, Breitwieser, 2015.

[13] Die Gesamtfläche ist die Summe der Flächen aller Leiterplatten, welche sich auf einem Warenträger befinden und somit gleichzeitig den ENIG-Prozess durchlaufen, Breitwieser, 2015.

[14] Die Kupferfläche ist die Summe der Kupferflächen (Aktivflächen) der Leiterplatten, welche sich auf einem Warenträger befinden und somit gleichzeitig den ENIG- Prozess durchlaufen, Breitwieser, 2015.

- Ätzreinigerkonzentration, Ätzrate und Rauigkeit wurden als weitere mögliche Parameter verworfen.
- Die entsprechenden Einheiten, in denen die Parameter gemessen wurden, sind in Tabelle 4.7 angeführt. Die Schichtstärken wurden in Mikrometer gemessen. Da die Goldschicht sehr dünn ist, wurde der betrachtete Mittelwert der Goldschichtstärke in Nanometer umgerechnet.

4.4.2 Datenqualität und Datenbereinigung

Mit dem Wissen über die Qualität des historischen Datensatzes wurde ein neuer Datensatz generiert. Dieser sollte unter anderem durch eine bessere Datenqualität zur Verbesserung der Modellierung der Schichtstärken beitragen. Allerdings wies der neue Datensatz wiederum Defizite auf und konnte nicht komplett verwendet werden. Die Kriterien der Datenqualität sind im Kapitel Datenqualität und Datenbereinigung, 4.1.2, erläutert. Die aufgetretenen Probleme werden in Abbildung 4.18 dargestellt.

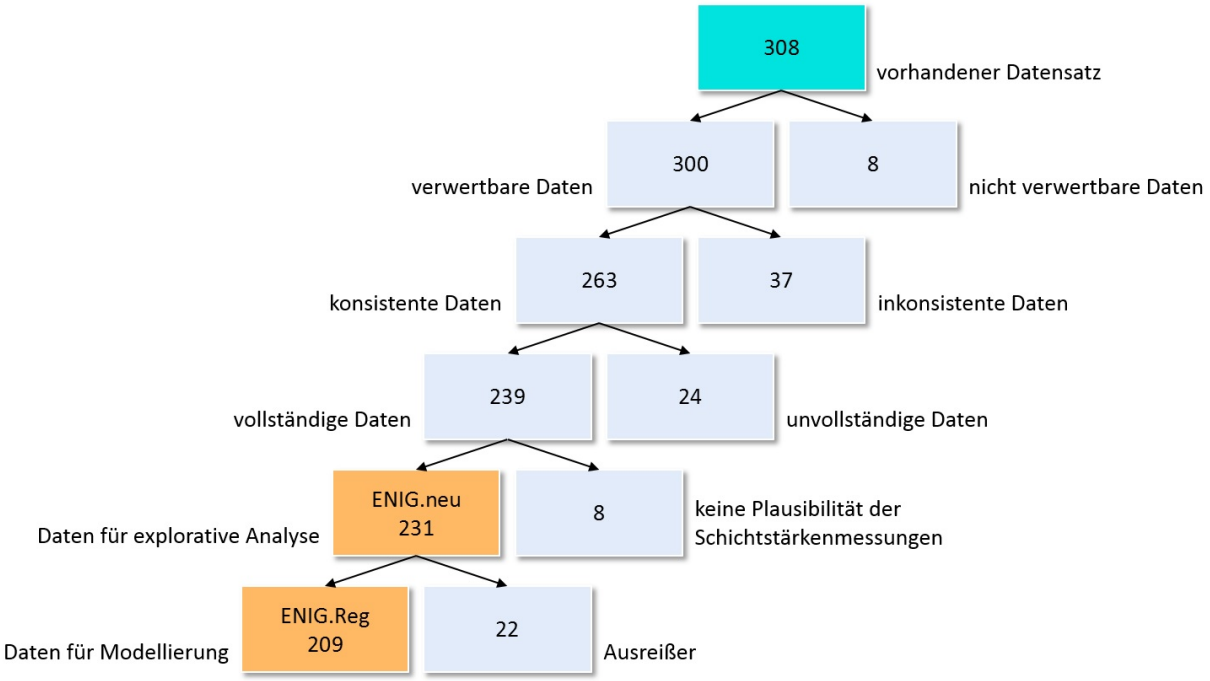

Abbildung 4.18: Modellierungsdatenbasis. Zwei Datensätze (ENIG.neu und ENIG.Reg) wurden für statistische Analysen herangezogen. Diese sind im Dateiformat CSV abgespeichert.

Im Zeitraum vom 26.11.2014 bis 28.02.2015 wurden insgesamt 308 Beobachtungen gemacht. Davon waren 300 Beobachtungen zur weiteren Verwendung geeignet. Die Daten sind in verschiedenen Datensätzen gespeichert, vgl. Tabelle 4.8. Da die Zusammenführung der unterschiedlichen Datensätze aufgrund mangelhafter Beschriftung nicht immer durchgeführt werden konnte, mussten 12% der verbleibenden Daten verworfen werden. Weitere 24 Beobachtungen waren nicht vollständig und daher für

eine mehrdimensionale Analyse nicht brauchbar. Zuletzt wurden acht Beobachtungen entfernt, bei denen die Messungen der Schichtstärken nicht plausibel erschienen.

Tabelle 4.8: Quellen der Variablen im Modellierungsdatensatz (ENIG.neu). Das größte Problem in der Erstellung des Datensatzes stellte die Zusammenführung der unterschiedlichen Teildatensätze dar.

Quelle	**Variable**	
Operator	Nummer	
	Datum	
	Schicht	
	Uhrzeit	
	MTO im Nickelbad	
Linie	Gesamtfläche	
	Aktivfläche	
	Minimaltemperatur aller Aktivbäder	
	Maximaltemperatur aller Aktivbäder	
	Liegezeit aller Aktivbäder	
Goldcontroller	MTO im Goldbad	
Chemielabor	Nickelbad	pH-Wert
		Konzentration
		Stabilitätswert
	Goldbad	pH-Wert
		Konzentration
		Ansatzlösung
		Kupfergehalt
		Nickelgehalt
Physiklabor	Nickelschichtstärken	
	Goldschichtstärken	

So blieben im Datensatz ENIG.neu 231 von 308 Beobachtungen für die explorative Analyse übrig, was einem Prozentsatz von 75% entspricht.

Ausreißer

Weiters sind Ausreißer im Datensatz vorhanden. Diese stellen eine Bedrohung dar, weil sie die Robustheit statistischer Verfahren gefährden können, vgl. Schendera, 2007. Im Fall der Regressionsanalyse können Ausreißer die Regressionskoeffizienten des Modells besonders beeinflussen und somit die Schätzer des Regressionsmodells verzerren. Für die Modellierung mit Hilfe von Regressionsmodellen wurden deshalb 22 Ausreißer entfernt. Der Datensatz, der für die Modellierung verwendet wurde, wurde mit ENIG.Reg gekennzeichnet.

Messfehler

Ein anderes Thema sind Messfehler. Daten, die sich aufgrund von Messfehlern deutlich von der Masse abheben (Ausreißer), können identifiziert und ausgemustert werden. Alle anderen zufälligen Messfehler können im Nachhinein nicht erkannt werden. Diese können nur durch Wiederholung ausfindig gemacht werden. In den verbliebenen Daten sind Messfehler enthalten. Da die Schichtstärken nur auf einem Messgerät gemessen wurden, besteht der Fehler ausschließlich aus Messfehlern des Messgeräts. Unterschiede in den Messungen bezüglich der Verwendung mehrerer Messgeräte sind also nicht gegeben.

Der Messfehler der Schichtstärkenmessungen wurde analysiert, siehe Riedler, 2015. Dafür wurden von je vier kleinen, mittleren und großen Pads die Schichtstärken fünfmal gemessen. Die gesamte Prozedur wurde dreimal wiederholt. Die Analyse ergab als Messfehler bei der Nickelschichtstärke eine Standardabweichung von 0.21 μm und bei der Goldschichtstärke eine Standardabweichung von 3.7 nm. Eine Messfehleranalyse der anderen Prozessparameter wurde nicht durchgeführt.

Für eine genaue Auswertung sollte zukünftig für jedes Messgerät eine Messfehleranalyse durchgeführt und nur ein Messgerät pro Parameter verwendet werden.

Timing

Einen weiterer Unsicherheitsfaktor stellt die zeitliche Übereinstimmung der Parameter einer Beobachtung dar. Das Timing aller Parameter pro Beobachtung sollte passen. Da die Probe für die chemischen Analysen (pH-Wert, Konzentration,. . . , siehe Tabelle 4.8) manuell gezogen werden, besteht die Möglichkeit, dass diese zeitlich nicht zu 100% mit den anderen Parametern, die automatisiert aufgezeichnet werden, zusammenpassen.

4.5 Explorative Analyse der Modellierungsdaten

Um den Datensatz statistisch zu analysieren, wurde eine explorative Datenanalyse durchgeführt, siehe Kapitel 3.1, Explorative Datenanalyse. Damit werden beobachtete Daten mit Hilfe von geeigneten Darstellungen und Kennzahlen untersucht. In der univariaten Analyse werden einzelne Variablen beschrieben und grafisch dargestellt, während in der multivariaten Analyse Zusammenhänge und Muster mehrerer Variablen identifiziert werden.

4.5.1 Univariate Analyse der Modellierungsdaten

Die erste Abbildung 4.19 zeigt alle gemessenen Gold- und Nickelschichtstärken im Zeitverlauf. Pro Beobachtung gibt es zwölf Werte der Goldschichtstärke und zwölf Werte der Nickelschichtstärke. Die unterschiedlichen Farben symbolisieren die unterschiedlichen Pad-Größen. Hier ist bereits zu erkennen, dass die Schichtstärken

auf den kleinen Pads größer sind als auf den mittleren Pads, und diese wiederum größer sind als auf den großen Pads. Bei beiden Grafiken ist zu sehen, dass der Spezifikationsbereich (Goldschichtstärke: $[0.06\ \mu m;\ 0.12\ \mu m]$ und Nickelschichtstärke: $[3\ \mu m;\ 6\ \mu m]$) bis auf wenige Ausreißer eingehalten wird. Beide Grafiken weisen auch keine bemerkenswerten zeitlichen Trends auf.

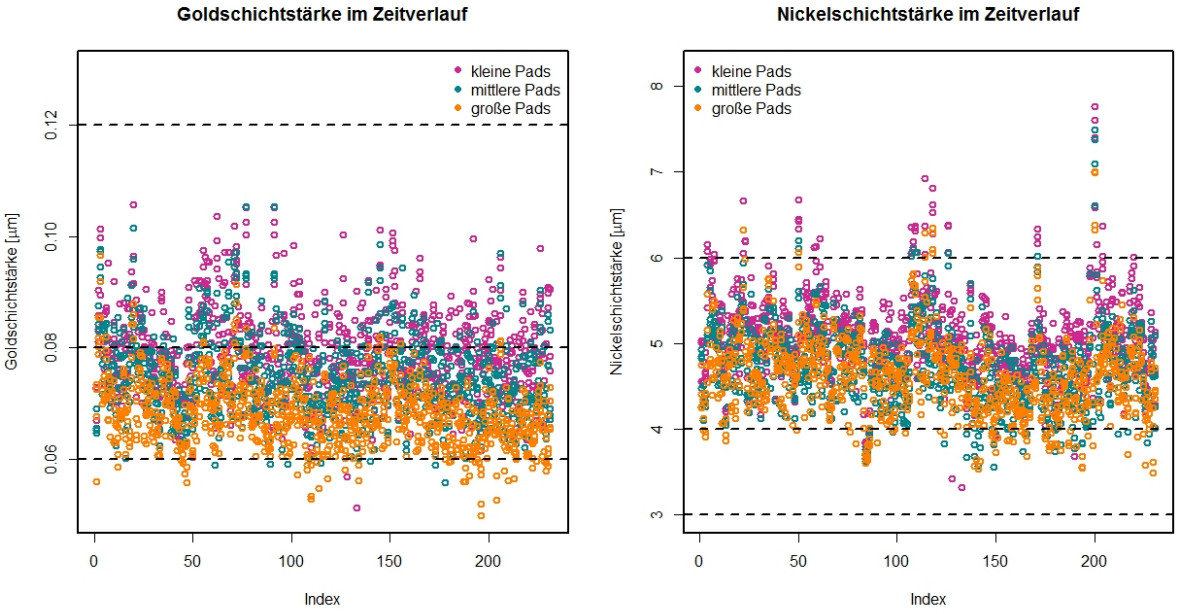

Abbildung 4.19: Goldschichtstärke und Nickelschichtstärke im Zeitverlauf im Modellierungsdatensatz (ENIG.neu). Die Schichtstärken unterscheiden sich je nach Pad-Größe.

Zur weiteren Analyse wurden die Messungen der Schichtstärken innerhalb der Pad-Größen gemittelt. Eine zeitliche Betrachtung dieser Messungen und die zugehörigen Boxplotserien zur Analyse der Verteilung liefern die Abbildungen 4.20 und 4.22.

Hierbei sind in Abbildung 4.20 zwei wesentliche Aspekte erkennbar. Erstens ist der Unterschied der mittleren Goldschichtstärken zwischen den Pad-Größen signifikant. Je kleiner die Pads, desto größer ist die Goldschichtstärke im Mittel. Zweitens bewegen sich die Schichtstärken der großen Pads ausschließlich im Idealbereich von 60 nm bis 80 nm und knapp darunter. Das bedeutet: je kleiner die Schichtstärke sein soll, desto größer sollte die Pad-Größe sein.

Bemerkung

- Die Goldschichtstärke ist auf großen Pads im Mittel kleiner als auf kleineren Pads. Aber aufgrund der größeren Fläche bei großen Pads muss eventuell mehr Gold für die gesamte Schutzschicht verwendet werden.

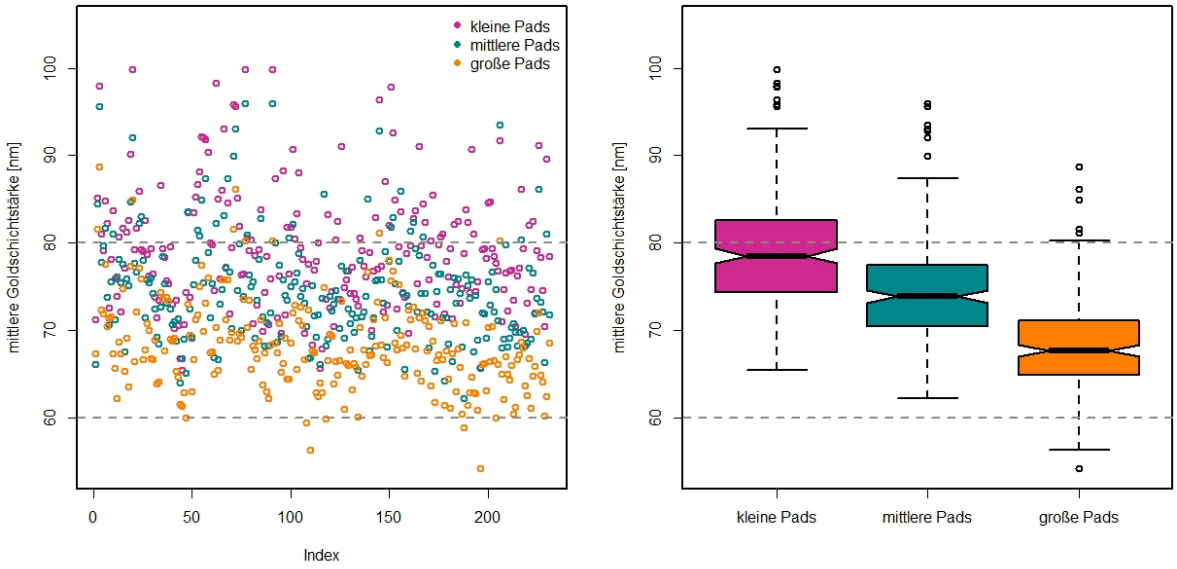

Abbildung 4.20: Mittlere Goldschichtstärke für den Modellierungsdatensatz (ENIG.neu) im Zeitverlauf. Je nach Pad-Größe ergeben sich verschiedene Verteilungen.

Eine andere Möglichkeit, Daten darzustellen, ist in Abbildung 4.21 ersichtlich. Sie zeigt drei Stängel-Blattdiagramme der mittleren Goldschichtstärke getrennt nach Pad-Größe. Die Verteilungen der Daten aller Pad-Größen sind annähernd symmetrisch und gleichen einer Normalverteilung.

```
The decimal point is at the |

  64 | 588
  66 | 89
  68 | 0416777
  70 | 00244467780345568
  72 | 00112256689011115556669999
  74 | 33344455667789334467889999
  76 | 12233455568889901223344679
  78 | 01133445556788911112222334444455667789
  80 | 0033356777000377899
  82 | 0013344556680345568888
  84 | 56678901113569
  86 | 116814
  88 | 1226
  90 | 247711278
  92 | 0261
  94 | 69
  96 | 38
  98 | 03888
```

```
The decimal point is at the |

  62 | 2
  64 | 0323557
  66 | 124578899667
  68 | 011122223355688000012222233448899
  70 | 000234477889123344555566789999
  72 | 000111122555578900112224456778999
  74 | 000122223466777800112234445566667889999
  76 | 00124445667779124557799
  78 | 001112569913789
  80 | 046889114578
  82 | 123480126
  84 | 57007
  86 | 0244
  88 |
  90 | 0
  92 | 0805
  94 | 6
  96 | 00
```

```
The decimal point is at the |

  54 | 3
  56 | 3
  58 | 959
  60 | 0012550022245568
  62 | 233456899999900113677899
  64 | 022235557788990000233456788 9
  66 | 011122333445566667777788889022333444555566778888889
  68 | 000333446666778889900011234446789
  70 | 004444445688891122333334455 99
  72 | 0012344779012445668
  74 | 4800378899
  76 | 1782456
  78 | 0
  80 | 333266
  82 |
  84 | 9
  86 | 2
  88 | 8
```

Abbildung 4.21: Stängel-Blattdiagrammserie der mittleren Goldschichtstärke getrennt nach Pad-Größe im Modellierungsdatensatz (ENIG.neu).

Wird die mittlere Nickelschichtstärke untersucht, ergibt sich ein ähnliches Ergebnis wie bei der mittleren Goldschichtstärke. Fast alle Beobachtungen sind im Spezifikationsbereich von 3 μm bis 6 μm - sogar im Idealbereich von 4 μm bis 6 μm. Jedoch ist der Unterschied der mittleren Nickelschichtstärke zwischen den mittleren und den großen Pads nicht so deutlich wie bei der mittleren Goldschichtstärke.

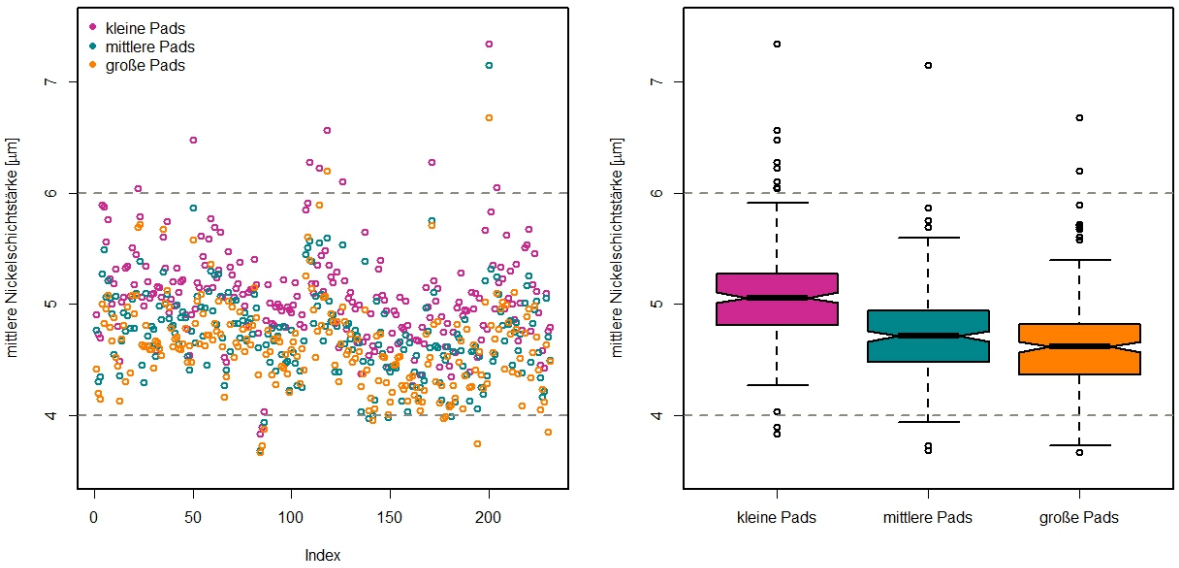

Abbildung 4.22: Mittlere Nickelschichtstärke für den Modellierungsdatensatz (ENIG.neu) im Zeitverlauf. Die Schichtstärken der mittleren und großen Pads unterscheiden sich nicht wesentlich.

Mittels Q-Q-Plot kann erkannt werden, ob die Daten annähernd normalverteilt sind. Abbildung 4.23 zeigt eine Q-Q-Plot-Serie der mittleren Nickelschichtstärke getrennt nach Pad-Größe. Die Schichtstärken scheinen normalverteilt zu sein. Ansonsten sind in jeder Grafik einige Ausreißer nach oben zu erkennen.

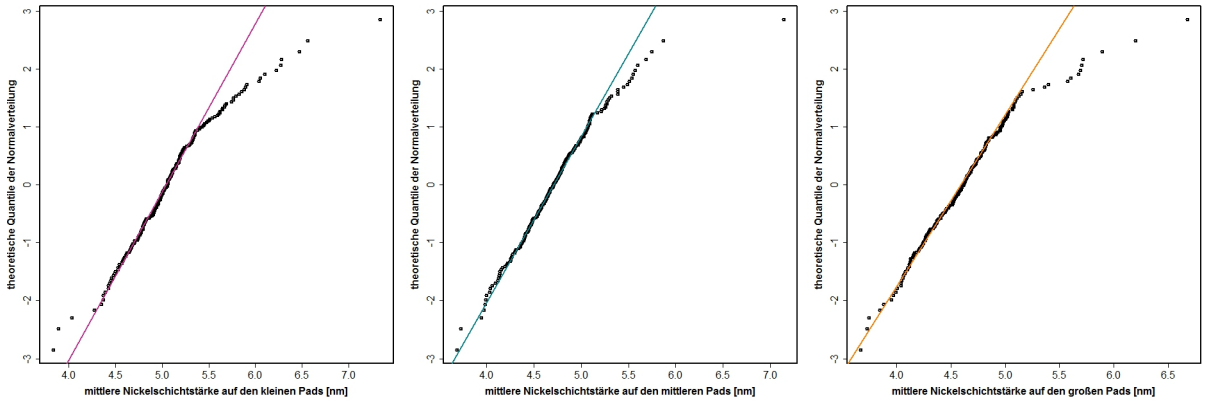

Abbildung 4.23: Q-Q-Plot-Serie der mittleren Nickelschichtstärke getrennt nach Pad-Größe im Modellierungsdatensatz (ENIG.neu).

Die Frage, ob nun auf jeder Testleiterplatte die Schichtstärken der kleinen Pads größer als die der mittleren Pads und die der mittleren Pads größer als die der großen Pads

sind, kann mit einer longitudinalen Betrachtung, Abbildung 4.24, veranschaulicht werden.

Auf der linken Seite ist die mittlere Goldschichtstärke abgebildet, während die rechte Grafik sich der mittleren Nickelschichtstärke widmet. Jede farbige Linie verbindet die Schichtstärken der kleinen, mittleren und großen Pads einer Testleiterplatte. Jede fallende Linie bestätigt die Behauptung, dass größere Pads kleinere Schichtstärken implizieren. Jedoch ist dies nicht auf allen Testleiterplatten der Fall. In 15% der Beobachtungen weisen die mittleren Pads eine größere Goldschichtstärke als die kleinen Pads auf und in 4% der Beobachtungen ist die Goldschichtstärke der großen Pads größer als die der mittleren Pads einer Testleiterplatte. In der rechten Grafik ist in 1% der Fälle die Nickelschichtstärke der kleinen Pads kleiner als die der mittleren Pads einer Testleiterplatte, jedoch ist in einem Drittel der Daten die Nickelschichtstärke der großen Pads größer als die der mittleren.

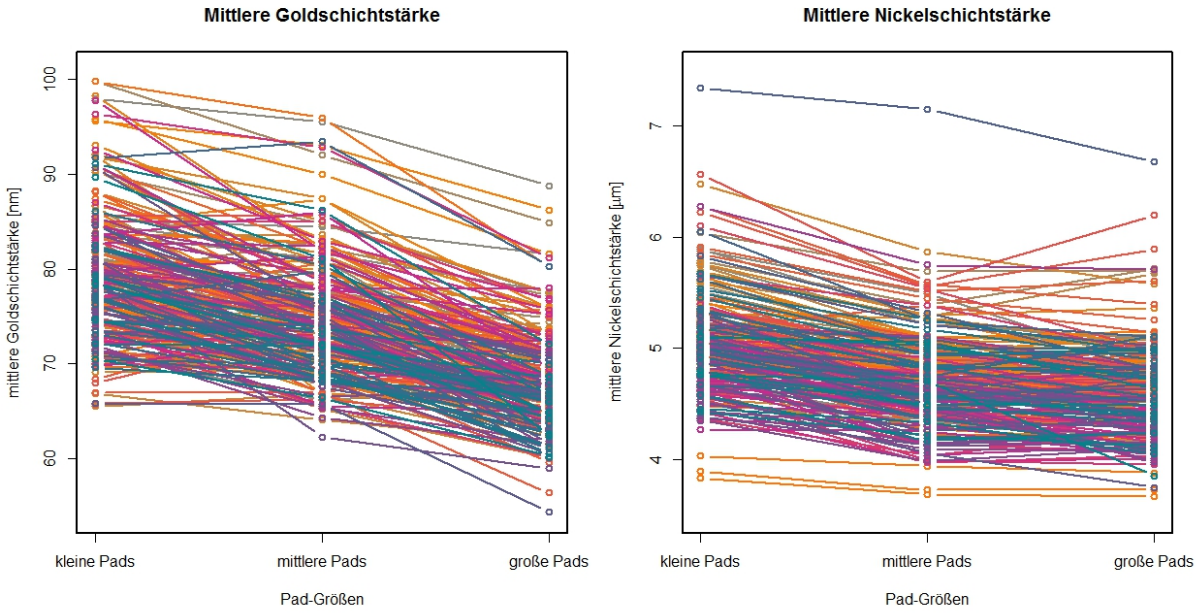

Abbildung 4.24: Longitudinalstudien der mittleren Goldschichtstärke und mittleren Nickelschichtstärke im Modellierungsdatensatz (ENIG.neu). Jede Linie verbindet die gemessenen Schichtstärken einer Testleiterplatte.

Einen Einblick in die Verteilung aller Variablen mittels statistischer Kenngrößen wie Minimum, Maximum, Mittelwert, Quantile und Standardabweichung liefert Tabelle 4.9. Dabei ist auffallend, dass einige Variablen nur gering schwanken. Dies ist darauf zurückzuführen, dass der Datensatz bei laufender Produktion aufgezeichnet wurde und somit keine außerplanmäßigen Einstellungen, welche die Produktion stören könnten, gemacht wurden.

Tabelle 4.9: Deskriptive Statistik der Prozessparameter und mittleren Schichtstärken im Modellierungsdatensatz (ENIG.neu). Die Kennzahlen sind: Minimum, 1. Quantil, Median, Mittelwert, 3. Quantil, Maximum, Standardabweichung.

	Min	$q_{0.25}$	$q_{0.5}$	MW	$q_{0.75}$	Max	Stdabw
Reiniger_Tmin	43.70	44.40	45.20	45.28	46.10	47.30	1.00
Reiniger_Tmax	43.90	44.70	45.60	45.55	46.30	47.50	0.98
Reiniger_Liegezeit	4.98	5.65	6.03	5.85	6.10	6.33	0.37
Microaetzen_Tmin	31.90	32.30	32.60	32.56	32.90	33.10	0.34
Microaetzen_Tmax	32.00	32.35	32.70	32.63	32.90	33.20	0.34
Microaetzen_Liegezeit	1.98	2.00	2.00	2.02	2.02	2.23	0.04
AuroDip_Tmin	66.70	68.80	69.70	69.70	70.50	71.30	0.94
AuroDip_Tmax	67.30	69.50	70.40	70.24	71.00	71.60	0.87
AuroDip_Liegezeit	4.98	5.03	5.37	5.47	5.98	6.35	0.46
AktivierenAu_Tmin	21.70	22.50	22.70	22.67	22.90	23.50	0.34
AktivierenAu_Tmax	21.70	22.50	22.70	22.74	23.00	23.60	0.36
AktivierenAu_Liegezeit	1.97	1.98	2.02	2.04	2.08	2.18	0.06
NipH	3.90	4.50	4.60	4.64	4.80	5.00	0.17
ChemNi_Tmin	77.00	80.30	80.40	80.39	80.60	81.20	0.41
ChemNi_Tmax	80.80	81.80	81.90	81.85	82.00	83.20	0.27
NiKonz	4.65	4.90	4.99	4.98	5.05	5.46	0.12
NiMTO	0.10	1.30	2.40	2.40	3.40	4.90	1.30
ChemNi_Liegezeit	23.00	23.13	23.22	23.52	23.33	30.27	1.10
NiStabi	3.00	6.00	6.00	5.90	6.00	8.00	0.69
AupH	5.50	5.70	5.70	5.71	5.70	5.80	0.05
SudGold_Tmin	81.10	82.60	83.50	83.34	84.10	84.90	0.93
SudGold_Tmax	81.90	83.70	84.40	84.35	85.10	85.80	0.89
Aukonz	0.51	0.55	0.56	0.56	0.57	0.78	0.03
AuMTO	0.09	4.48	7.97	8.16	11.61	16.32	4.30
SudGold_Liegezeit	11.07	13.07	13.12	13.50	13.23	16.35	1.04
AuAnsatz	93.21	104.00	106.40	106.20	108.90	114.60	3.54
AuCu	0.34	3.46	3.92	4.04	4.50	10.54	1.03
AuNi	53.35	335.60	534.60	497.90	665.60	829.90	199.28
Heissspuele_Tmin	54.80	68.30	68.70	68.68	69.20	69.60	1.09
Heissspuele_Tmax	66.30	68.60	69.10	69.07	69.60	70.00	0.60
Heissspuele_Liegezeit	1.58	2.37	2.52	2.48	2.58	2.75	0.12
Trockner_Tmin	71.30	72.60	72.90	72.81	73.20	73.40	0.45
Trockner_Tmax	76.10	76.20	76.30	76.32	76.40	76.70	0.11
Trockner_Liegezeit	12.92	14.63	14.67	14.70	14.68	17.62	0.33
Total_Area	185.80	712.70	816.00	796.50	913.20	1264.00	169.07
Copper_Area	110.70	143.90	176.60	187.20	219.30	473.70	54.49
MWNiK	3.83	4.81	5.06	5.08	5.27	7.34	0.44

Fortsetzung nächste Seite

Tabelle 4.9 – *Fortsetzung von vorheriger Seite*

	Min	$q_{0.25}$	$q_{0.5}$	MW	$q_{0.75}$	Max	Stdabw
MWNiM	3.68	4.47	4.71	4.72	4.95	7.14	0.41
MWNiG	3.67	4.37	4.62	4.62	4.82	6.68	0.41
MWAuK	65.47	74.34	78.54	79.18	82.59	99.82	6.93
MWAuM	62.23	70.43	73.91	74.67	77.48	95.98	6.13
MWAuG	54.29	64.93	67.72	68.28	71.14	88.77	5.21

4.5.2 Multivariate Analyse der Modellierungsdaten

Als weiteren Schritt in der EDA werden Zusammenhänge der Parameter untersucht. Abbildung 4.25 gibt einen Überblick über mögliche Zusammenhänge der Variablen.

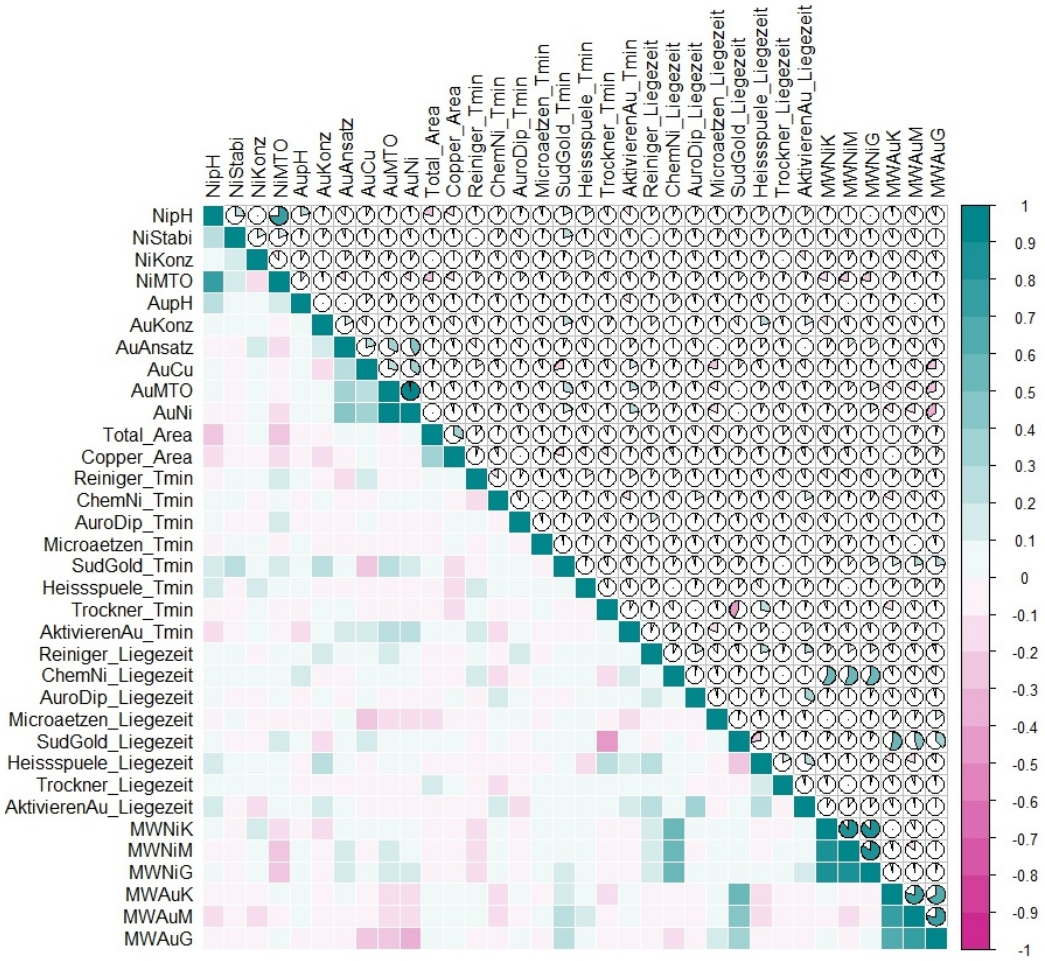

Abbildung 4.25: Korrelationsmatrix der Variablen im Modellierungsdatensatz (ENIG.neu). Türkis weist auf eine hohe positive Korrelation hin, weiß auf keine Korrelation und pink auf eine hohe negative Korrelation.

Die Maximaltemperaturen der Aktivbäder wurden zur besseren Lesbarkeit nicht abgebildet. Auch im neuen Datensatz sind die Zusammenhänge im Großen und Ganzen sehr schwach. Das spiegelt sich einerseits auf der linken unteren Hälfte in den blassen Farben wieder, wie auch in der rechten oberen Hälfte bei den Kreisen.

Die Korrelationskoeffizienten zwischen den mittleren Goldschichtstärken und den Prozessparametern sind in Tabelle 4.10 aufgelistet. Hierbei ist nochmal erkennbar, dass die Zusammenhänge sehr gering sind. Den stärksten Zusammenhang mit 0.37 bis 0.53 (je nach Pad-Größe) hat die mittlere Goldschichtstärke mit der Liegezeit im Goldbad. Alle anderen Korrelationskoeffizienten bewegen sich im Intervall $[-0.34;\ 0.24]$.

Tabelle 4.10: Korrelationskoeffizienten zwischen den Goldschichtstärken und den Prozessparametern im Modellierungsdatensatz (ENIG.neu).

	MWAuK	MWAuM	MWAuG
SudGold_Liegezeit	0.53	0.43	0.37
SudGold_Tmin	0.18	0.20	0.23
SudGold_Tmax	0.18	0.21	0.24
AktivierenAu_Tmax	0.12	0.09	0.01
Heissspuele_Tmax	0.10	0.14	0.13
AktivierenAu_Tmin	0.09	0.07	-0.01
Microaetzen_Liegezeit	0.08	0.07	0.14
Reiniger_Tmin	0.07	0.09	-0.05
Heissspuele_Tmin	0.07	0.12	0.07
NiMTO	0.06	-0.02	0.01
AupH	0.05	-0.00	-0.03
AuAnsatz	0.05	-0.10	-0.06
Reiniger_Tmax	0.05	0.07	-0.07
AuroDip_Tmin	0.01	-0.02	0.01
Total_Area	0.01	0.06	0.04
Copper_Area	0.00	0.09	0.06
AuroDip_Tmax	0.00	-0.03	0.00
MWNiK	0.00	-0.08	-0.00
Microaetzen_Tmin	-0.01	0.00	-0.04
Microaetzen_Tmax	-0.01	0.01	-0.04
AuCu	-0.01	-0.10	-0.25
NipH	-0.02	-0.11	-0.09
MWNiG	-0.03	-0.02	0.03
MWNiM	-0.04	-0.14	-0.01
AktivierenAu_Liegezeit	-0.05	-0.02	-0.01
ChemNi_Liegezeit	-0.06	-0.09	-0.11
Trockner_Liegezeit	-0.06	-0.05	-0.03
NiKonz	-0.07	-0.13	-0.08
NiStabi	-0.07	-0.08	-0.01
AuroDip_Liegezeit	-0.08	-0.07	-0.04
Trockner_Tmax	-0.09	-0.12	-0.17
Reiniger_Liegezeit	-0.09	-0.06	-0.06
AuKonz	-0.10	-0.07	-0.02
AuMTO	-0.14	-0.17	-0.29
AuNi	-0.15	-0.20	-0.34
ChemNi_Tmin	-0.16	-0.11	-0.10
Heissspuele_Liegezeit	-0.17	-0.19	-0.10
ChemNi_Tmax	-0.20	-0.22	-0.26
Trockner_Tmin	-0.20	-0.07	-0.03

Der Grund für den schwachen linearen Zusammenhang ist einfach erkennbar, wenn die zugehörigen Scatterplots betrachtet werden. Ein Beispiel zeigt Abbildung 4.26. Darauf sind die Liegezeit und die Minimaltemperatur sowie der Nickelgehalt im Goldbad und die mittlere Goldschichtstärke auf den kleinen Pads abgebildet.

Die Liegezeit und die Minimaltemperatur sind die Variablen mit dem stärksten positiven Zusammenhang mit der mittleren Goldschichtstärke der kleinen Pads. Jedoch weisen die mittleren Goldschichtstärken bei einer Liegezeiten von rund 13 bzw. rund 16 Minuten eine große Schwankungsbreite auf. Da die Beobachtungen vorwiegend bei einer Liegezeit von rund 13 Minuten bzw. 16 Minuten gemacht wurden, muss mit konkreten Aussagen vorsichtig umgegangen werden.

Eine andere Situation zeigen die Scatterplots zwischen der Minimaltemperatur im Goldbad bzw. des Nickelgehalts im Goldbad mit der Goldschichtstärke. Hier sind vor allem Punktwolken erkennbar, die auf keinen signifikanten Zusammenhang hinweisen.

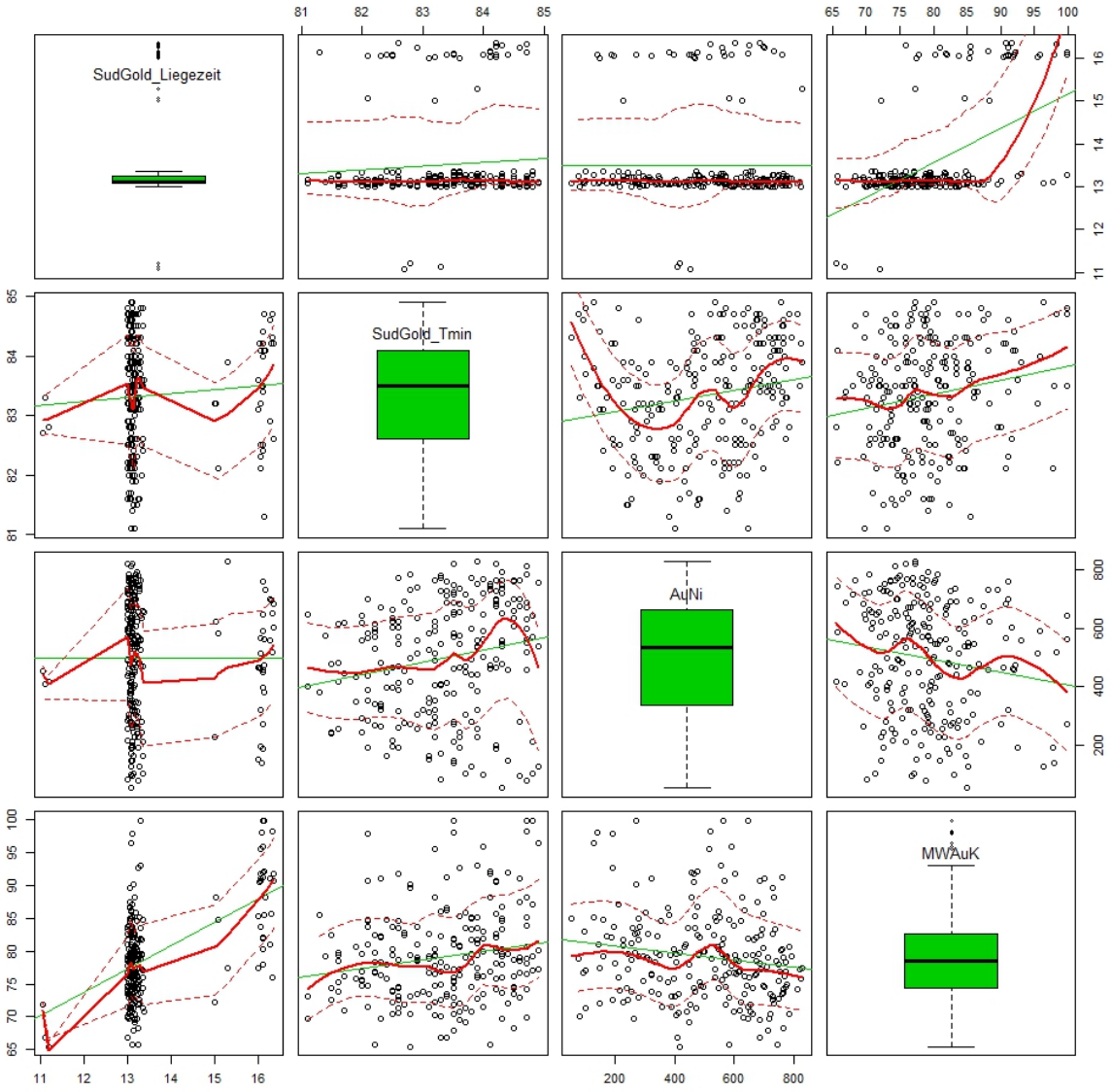

Abbildung 4.26: Scatterplotmatrix einiger Variablen im Modellierungsdatensatz (ENIG.neu). Statt signifikanter Zusammenhänge sind vorwiegend Punktwolken ohne Struktur zu erkennen.

Eine weitere Darstellungsmöglichkeit der Zusammenhänge ist durch Beeswarm-Boxplots gegeben. Ein Beispiel dafür zeigt Abbildung 4.27. Die Boxplots stellen die Verteilung der mittleren Goldschichtstärke der kleinen Pads dar. Die Bienenschwärme sind anhand eines Prozessparameters eingefärbt. In den zwei oberen Grafiken dominieren die Farben grün bzw. rot, da die meisten Beobachtungen im Goldbad einen pH-Wert von 5.7 und eine Konzentration größer als 0.54 aber kleiner gleich 0.58 aufweisen. Die restlichen Farben entsprechen den Klassen wie in den jeweiligen Legenden ersichtlich. Die verschiedenen Farben zeigen kein Muster, die auf spezielle Zusammenhänge hinweisen.

In den zwei unteren Grafiken dominieren keine speziellen Farben. Alle Farben sind

vertreten. Wiederum ist kein Muster sichtbar. Das heißt kleine bzw. große Goldschichtstärken wurden sowohl bei einer konzentrierten als auch bei einer weniger konzentrierten Ansatzlösung gemessen. Ebenso wurden kleine bzw. große Goldschichtstärken sowohl bei kleinen als auch bei großen MTO-Werten des Goldbads beobachtet. Abbildung 4.27 spiegelt die Struktur, wie oben beobachtet, wider.

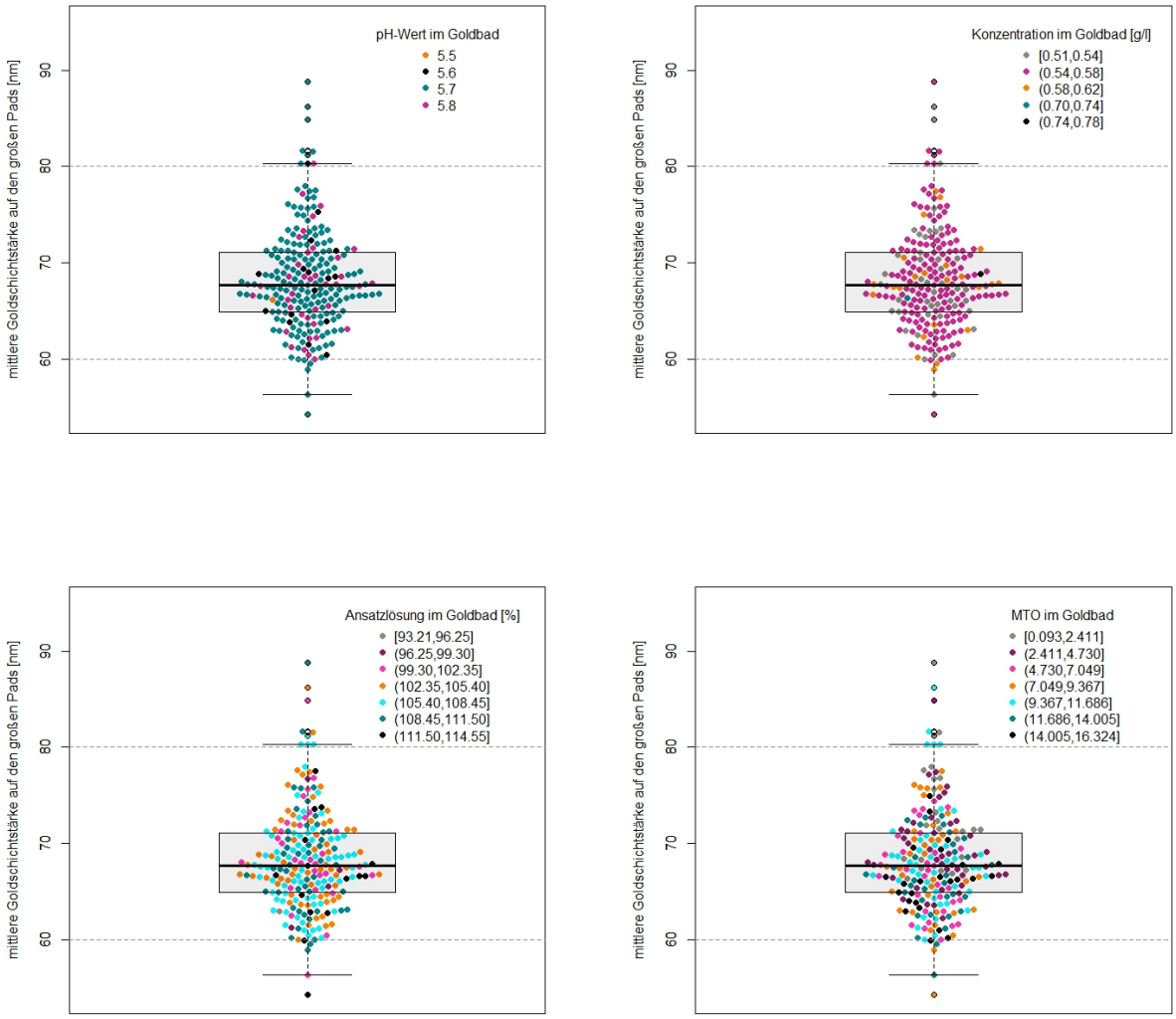

Abbildung 4.27: Beeswarm-Boxplots zur bivariaten Analyse im Modellierungsdatensatz (ENIG.neu). Dabei werden Punkte im Boxplot der mittleren Goldschichtstärke der großen Pads anhand verschiedener Prozessparameter kategorisch eingefärbt.

Um Zusammenhänge zwischen mehreren Parametern zu untersuchen und grafisch darzustellen eignen sich Scatterplot zwischen zwei Variablen bei denen eine dritte, vierte, fünfte... Variable durch Farben, Formen oder Größen symbolisiert wird. Ein Beispiel hierfür zeigt Abbildung 4.28. Die zwei Scatterplots zeigen den Zusammenhang zwischen der Liegezeit im Goldbad auf der linken Seite und der Minimaltemperatur

im Goldbad auf der rechten Seite mit der mittleren Goldschichtstärke der kleinen Pads. Die unterschiedlichen Farben stellen die verschiedenen Klassen des Nickelgehalts im Goldbad dar.

Wie auch schon in der bivariaten Analyse erkennbar, sind kaum Zusammenhänge erkennbar. Dies kommt auch in dieser Abbildung zum Ausdruck. Es scheint, als ob die Farben zufällig verteilt sind. Jegliche Kombinationen der Variablen im Modellierungsdatensatz liefern eher zufällige Muster.

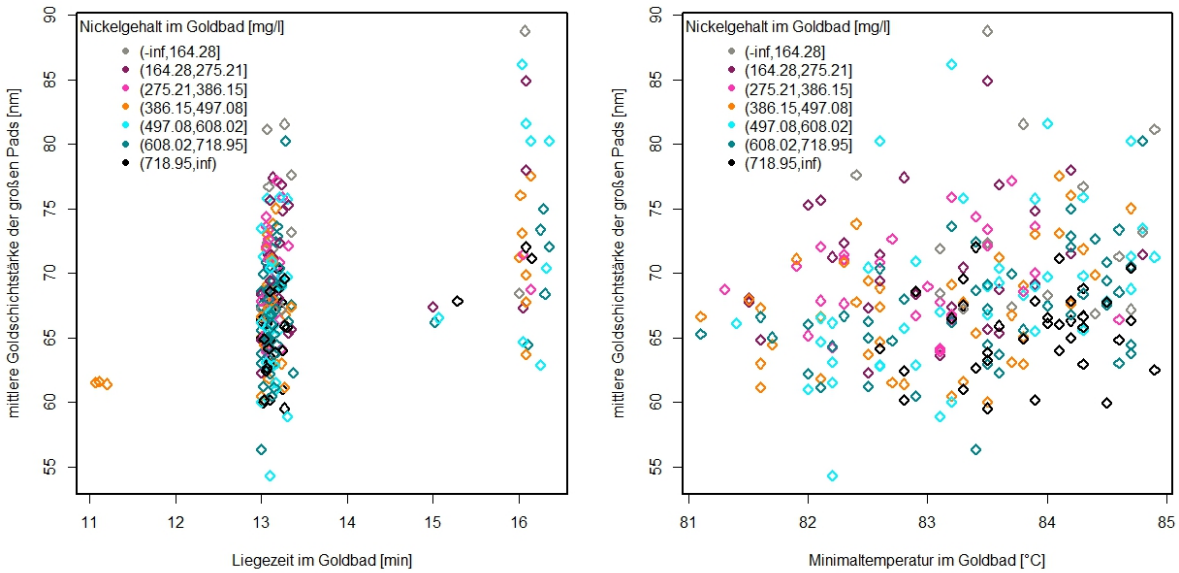

Abbildung 4.28: Scatterplots: Liegezeit im Goldbad und mittlere Goldschichtstärke der kleinen Pads inklusive Nickelgehalt im Goldbad bzw. Minimaltemperatur im Goldbad und mittlere Goldschichtstärke der kleinen Pads inklusive Nickelgehalt im Goldbad (ENIG.neu).

Eine andere Möglichkeit der Analyse und Aufspaltung der Daten bietet die Variable Uhrzeit. Zu allen unterschiedlichen Uhrzeiten sind, wie in Abbildung 4.29 dargestellt, ungefähr gleich viele Beobachtungen gemacht worden. Eine Möglichkeit zur grafischen Veranschaulichung, ob es Unterschiede in den Schichtstärken zu den verschiedenen Uhrzeiten gibt, bieten Boxplotserien.

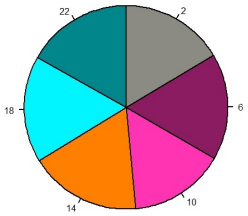

Abbildung 4.29: Zeitliche Verteilung der Beobachtungen im Modellierungsdatensatz (ENIG.neu). Die Beobachtungen sind zeitlich annähernd gleichverteilt.

Abbildung 4.30 und Abbildung 4.31 stellen die mittlere Goldschichtstärke bzw. die mittlere Nickelschichtstärke getrennt nach Pad-Größe und Uhrzeit dar.

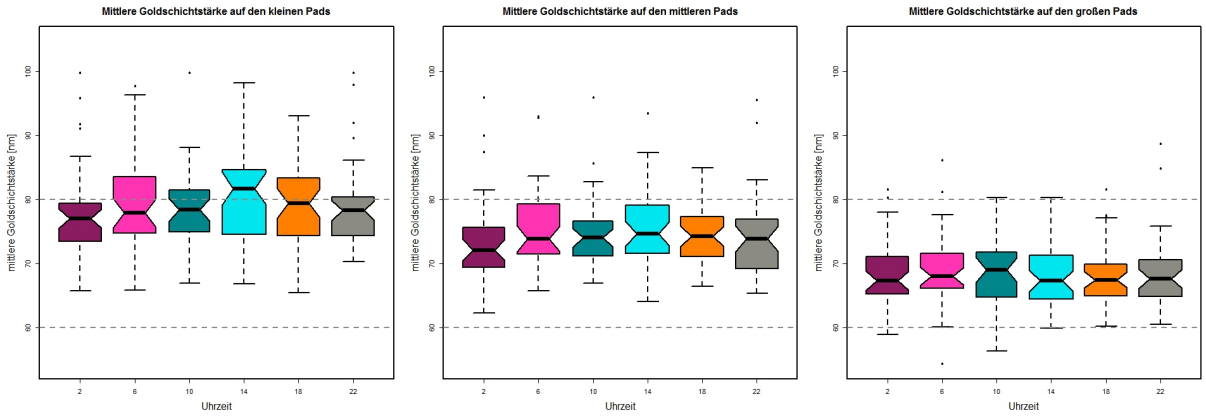

Abbildung 4.30: Boxplotserie der mittleren Goldschichtstärke getrennt nach Uhrzeit im Modellierungsdatensatz (ENIG.neu).

Bei beiden Abbildungen sind keine Besonderheiten erkennbar. Leichte Schwankungen in den interquantilen Bereichen und der Länge der Tails der verschiedenen Boxplots sind zu erkennen, aber kein durchgehendes Muster, das sich über alle Grafiken zieht.

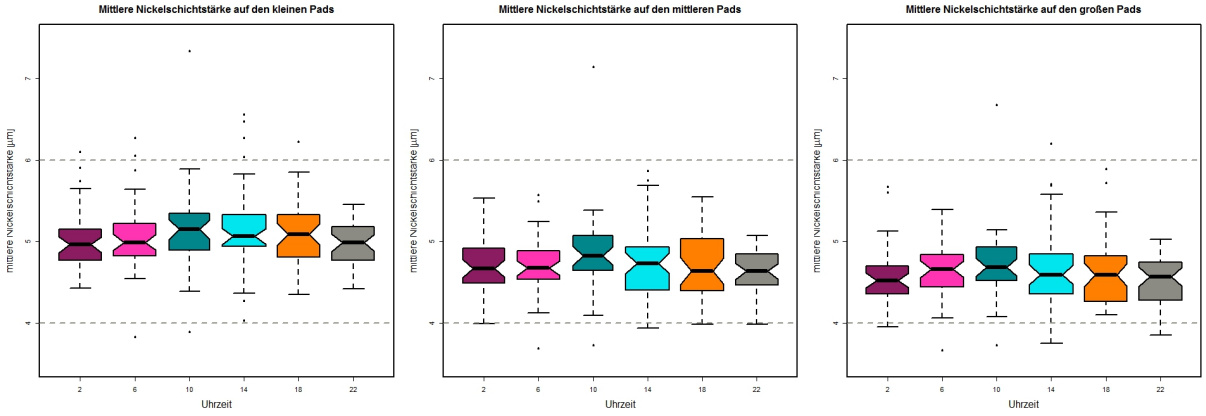

Abbildung 4.31: Boxplotserie der mittleren Nickelschichtstärke getrennt nach Uhrzeit im Modellierungsdatensatz (ENIG.neu).

4.6 Regressionsanalyse der Modellierungsdaten

Mittels linearer Regressionsmodelle soll die Beziehung zwischen der mittleren Goldschichtstärke der unterschiedlichen Pad-Größen und den Prozessparametern beschrieben werden. Je mehr Prädiktoren in einem Modell enthalten sind, desto sensitiver ist

das Modell. Soll das Modell nur zur Beschreibung eines spezifischen Datensatzes verwendet werden, spielt die Anzahl der Prädiktoren keine große Rolle. Soll das Modell aber, wie in diesem Fall, vorwiegend zur Vorhersage verwendet werden, sollte das System gröber sein, dafür aber stabil, vgl. Stadlober, 2008.

4.6.1 Multivariates Modell für die mittlere Goldschichtstärke

Der Datensatz beinhaltet drei unterschiedliche Zielvariablen: die mittleren Goldschichtstärken der drei verschiedenen Pad-Größen. Für die Modellierung wurden multivariate multiple Regressionsmodelle verwendet. Das Ziel dabei ist, alle drei Zielvariablen gleichzeitig mit demselben Satz an Prädiktoren zu modellieren, siehe Kapitel 3.2, Lineare Regression.

Zuerst wurde für jede Zielvariable ein Regressionsmodell ausgearbeitet. 36 Prozessparameter standen als mögliche Prädiktoren zu Verfügung. Die Nickelschichtstärke wurde nicht in den Pool der Prädiktoren aufgenommen, da sie ein Produkt des ENIG-Prozesses ist und damit nicht kontrollierbar.

Falls jede Variable einfach im Modell enthalten sein kann, ohne zusätzlich Transformationen und Interaktionen zu berücksichtigen, stehen $2^{36} = 68\,719\,476\,736$ mögliche Regressionsmodelle zur Beschreibung jeder der drei Zielvariablen zu Verfügung. Um nicht alle Modelle einzeln zu überprüfen, wurden für die Prozedur der Variablenselektion die Ergebnisse der explorativen Analyse verwendet und die **R**-Funktion *regsubsets* aus dem Paket *leaps* für eine vollständige Modellselektion basierend auf den Modellwahlkriterien BIC, C_p Kriterium und R^2_{adj} bzw. die **R**-Funktion *step* für eine schrittweise Selektion basierend auf Akaike´s Informationskriterium angewandt, siehe Kapitel 3.6, Modellbildung. Genauere Informationen zur Verwendung der **R**-Funktionen *regsubsets* und *step* finden sich in Lumley, 2015, und dem Hilfe-Handbuch von **R**.

Anschließend wurde anhand der ausgewählten multiplen Regressionsmodelle ein multivariates Regressionsmodell zur Beschreibung der mittleren Goldschichtstärke kombiniert.

Die Modellfindungsprozedur lässt sich als Kreislauf darstellen. Startpunkt ist die Suche nach der passenden Teilgruppe an Prädiktoren. Diese wird mit Hilfe statistischer Kennwerte, des Bayesschen Informationskriteriums (BIC), Mallows-C_p-Statistik und des adjustierten Bestimmtheitsmaßes (R^2_{adj}) gesucht. Dabei gilt es, das Informationskriterium zu minimieren und das Bestimmtheitsmaß zu maximieren. Der C_p-Wert hingegen sollte möglichst gleich groß wie die Anzahl an Prädiktoren + 1 sein. Anschließend wird das Modell mittels Grafiken überprüft und entsprechend angepasst. Um wiederum die Menge an Prädiktoren für das modifizierte Modell zu überprüfen, beginnt die Prozedur von vorne, vgl. Abbildung 4.32.

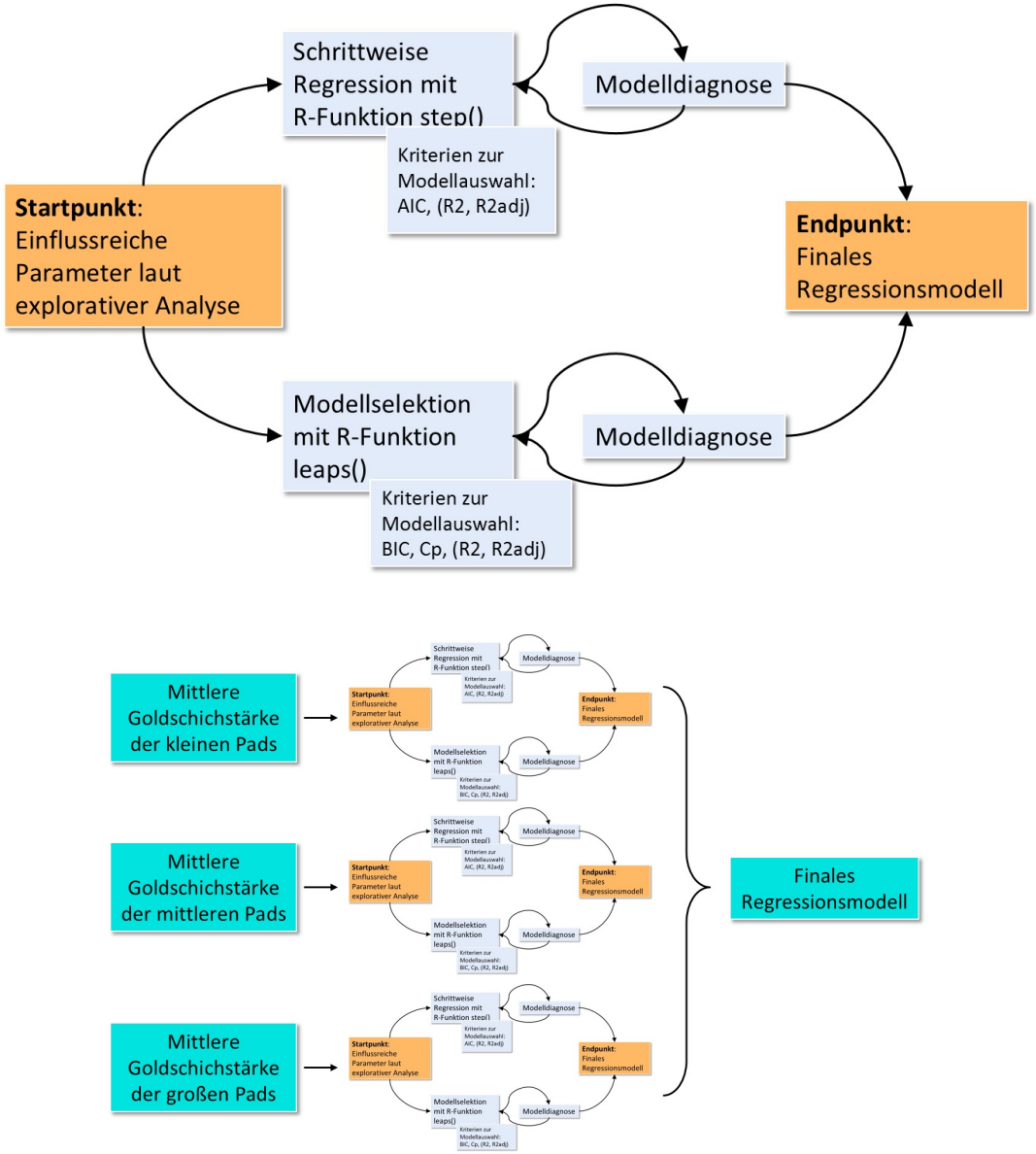

Abbildung 4.32: Prozess der Modellselektion. Die Auswahl eines Modells kann als Kreislauf dargestellt werden. Sie erfolgt einerseits durch eine schrittweise Regression und basiert andererseits auf speziellen mathematischen Kriterien.

Das Ergebnis der multivariaten multiplen Regression liefert Tabelle 4.11. Die Gruppe der Prädiktoren besteht aus fünf Variablen:

- der Liegezeit im Goldbad,
- dem Nickelgehalt im Goldbad,
- der Minimaltemperatur im Goldbad,
- der Maximaltemperatur im Nickelbad und
- der Maximaltemperatur im Aktivator.

Die Hinzunahme von Wechselwirkungen und Transformationen der Prädiktoren in das multivariate Modell ergab keine allgemeine Verbesserung der Modellgüte.

Tabelle 4.11: Multivariates Regressionsmodell für die mittlere Goldschichtstärke im reduzierten Modellierungsdatensatz (ENIG.Reg).

$MWAuK^{-1}$, $MWAuM^{-1}$, $MWAuG^{-1} \sim$ SudGold_Liegezeit+AuNi+SudGold_Tmin+ChemNi_Tmax+AktivierenAu_Tmax			
Coefficients:			
	$MWAuK^{-1}$	$MWAuM^{-1}$	$MWAuG^{-1}$
(Intercept)	-2.561e-02	-2.592e-02	-3.692e-02
SudGold_Liegezeit	-4.546e-04	-3.387e-04	-2.895e-04
AuNi	1.169e-06	1.346e-06	2.172e-06
SudGold_Tmin	-2.217e-04	-2.658e-04	-3.580e-04
ChemNi_Tmax	8.872e-04	9.033e-04	1.116e-03
AktivierenAu_Tmax	-4.512e-04	-3.722e-04	-3.080e-04

Alle drei Zielvariablen wurden aufgrund der Diagnoseplots invertiert. Da alle drei Zielvariablen auf gleiche Art transformiert wurden, ist ein Vergleich der Regressionskoeffizienten der drei Zielvariablen möglich. Die Regressionskoeffizienten jedes Prädiktors haben die gleichen Vorzeichen und sind ähnlich groß. Die Regressionskoeffizienten verschiedener Prädiktoren sind nicht vergleichbar, da die Prädiktorvariablen unterschiedliche Größen haben. Dazu müssten die standardisierten Regressionskoeffizienten herangezogen werden.

Eine detailliertere Betrachtung der verschiedenen Modelle liefern die nächsten Tabellen und Grafiken.

4.6.2 Modell für die mittlere Goldschichtstärke der kleinen Pads

Den Beginn macht das Modell für die mittlere Goldschichtstärke der kleinen Pads, vgl. Tabelle 4.12. Die Variablen Liegezeit und Nickelgehalt im Goldbad und die Maximaltemperatur im Nickelbad sind hoch signifikant. Die Minimaltemperatur im Goldbad ist etwas weniger signifikant und die Maximaltemperatur im Aktivator hat von den Prädiktoren den größten p-Wert mit 0.0101. Das Bestimmtheitsmaß und das adjustierte Bestimmtheitsmaß sind mit 36.2% und 34.7% fast gleich groß. Demnach sind keine redundanten Variablen enthalten. Die F-Statistik zeigt an, dass die Hypothese: $\beta_i = 0,\ \forall i$, verworfen werden kann.

Aufgrund der Ergebnisse der explorativen Analyse der Daten, welche die Schwankungen der Goldschichtstärke bei fester Parametereinstellung und die schwachen Zusammenhänge der Prozessparameter sichtbar machte, ist das Ergebnis des Regressionsmodells im Hinblick auf Modellgüte und Modellfehler durchaus nachvollziehbar.

Die Diagnoseplots, Abbildung 4.33 weisen keine besonderen Auffälligkeiten auf.

Tabelle 4.12: Lineares Regressionsmodell für die mittlere Goldschichtstärke der kleinen Pads im reduzierten Modellierungsdatensatz (ENIG.Reg).

$MWAuK^{-1} \sim$ SudGold_Liegezeit+AuNi+SudGold_Tmin+ChemNi_Tmax+AktivierenAu_Tmax					
	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	-2.561e-02	2.193e-02	-1.168	0.244161	
SudGold_Liegezeit	-4.546e-04	5.745e-05	-7.913	1.6e-13	***
AuNi	1.169e-06	3.195e-07	3.659	0.000322	***
SudGold_Tmin	-2.217e-04	6.706e-05	-3.306	0.001118	**
ChemNi_Tmax	8.872e-04	2.449e-04	3.623	0.000368	***
AktivierenAu_Tmax	-4.512e-04	1.739e-04	-2.595	0.010148	*
Observations	209				
R^2	0.362				
Adjusted R^2	0.347				
Residual Std. Error	0.001 (df = 203)				
F Statistic	23.078 *** (df = 5; 203)				
Signif. codes:	*** p<0.001 ** p<0.01 * p<0.05 . p<0.1				

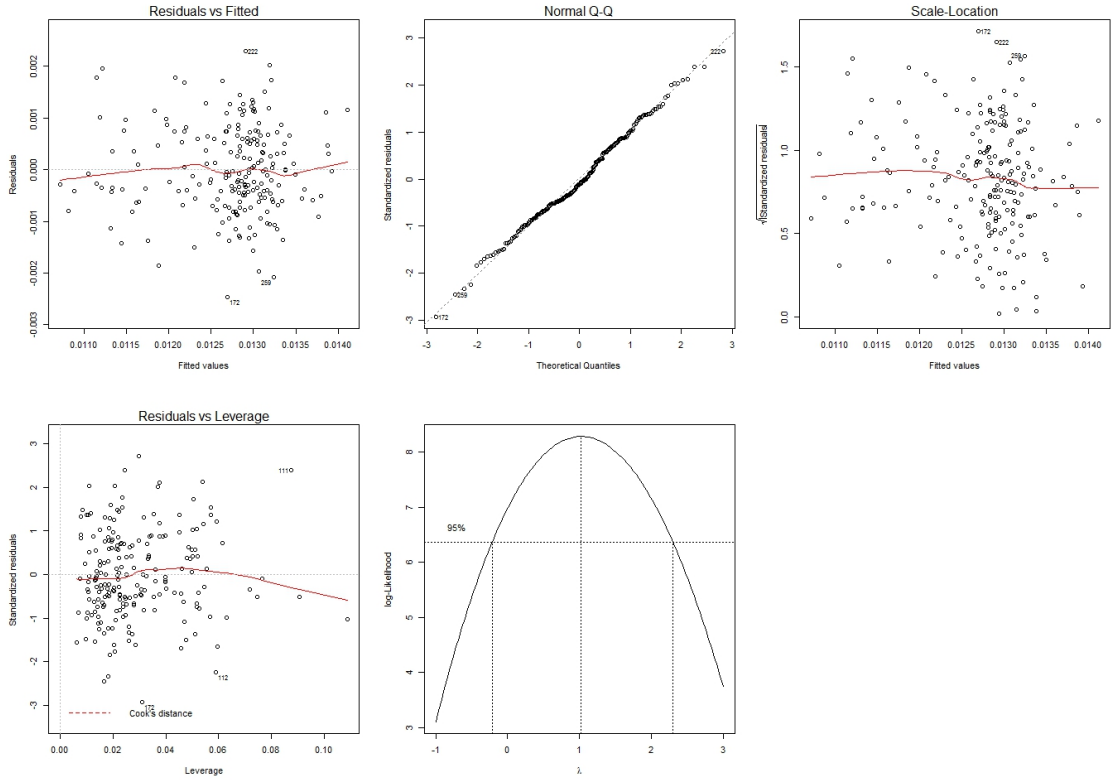

Abbildung 4.33: Modelldiagnoseplots des linearen Regressionsmodells für die mittlere Goldschichtstärke der kleinen Pads im reduzierten Modellierungsdatensatz (ENIG.Reg). Mittels grafischer Methoden können die Modellannahmen überprüft werden.

Der erste Plot zeigt einen Scatterplot der Residuen gegen die gefitteten Werte. Dabei ist kein Trend der Residuen zu entdecken. Die zweite Grafik ist ein Q-Q-Plot. Hier werden die standardisierten Residuen gegen die theoretische Quantile der Normalverteilung aufgetragen. Damit soll die Grafik über die Normalverteilung der Residuen informieren. Auch diese Grafik zeigt keine Besonderheiten, da die Punkte sehr gut auf der Referenzlinie liegen. Es liegt demnach kein Widerspruch zur Annahme der Normalverteilung vor. Die dritte Grafik dient der Überprüfung der Varianzhomogenität. Dabei werden die Wurzeln der standardisierten Residuen gegen die gefitteten Werte abgebildet. Es sind keine Auffälligkeiten sichtbar, sodass von der Gleichheit der Varianzen ausgegangen werden kann. Der nächste Plot verifiziert den Einfluss einzelner Beobachtungen. Die Regressionskoeffizienten werden von keinen Beobachtungen besonders verzerrt. Die letzte Grafik überprüft, ob eine Transformation der Zielvariablen notwendig ist. In diesem Fall ist keine weitere Transformation notwendig.

Zur grafischen Darstellung des Regressionsmodells wurde ein Regressionsbaum gezeichnet, Abbildung 4.34.

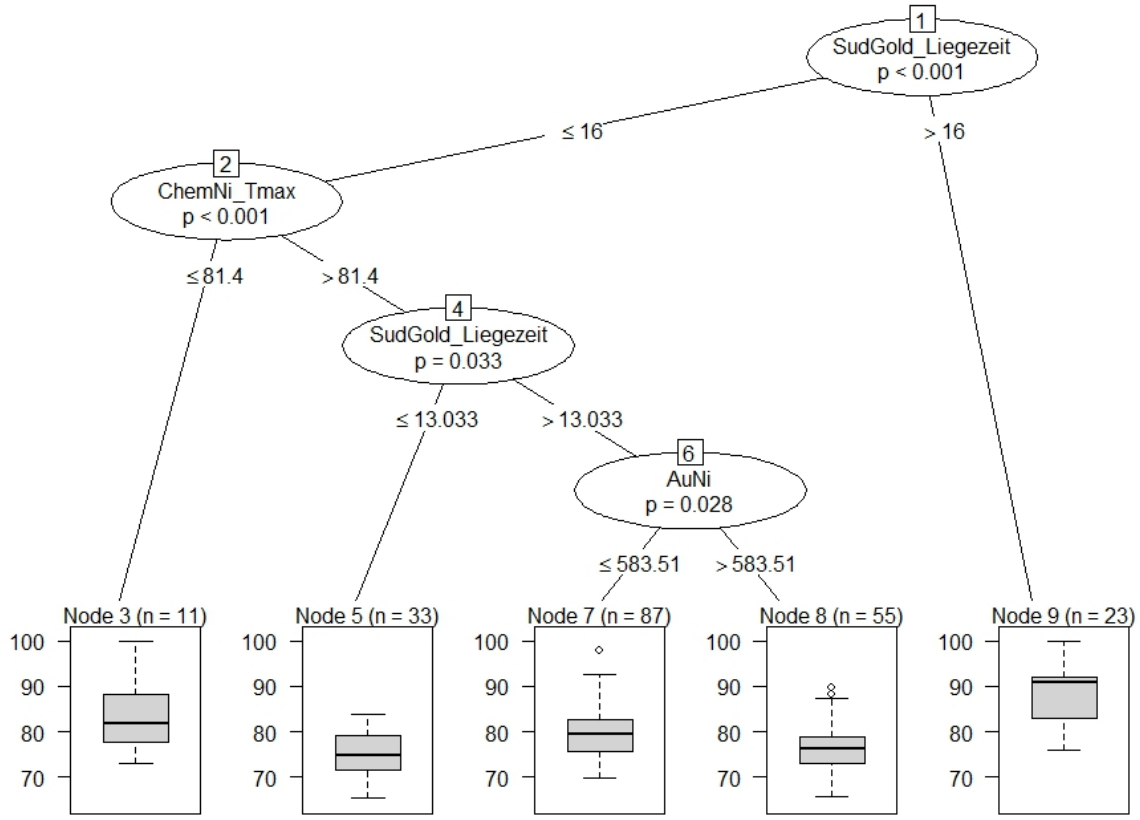

Abbildung 4.34: Regressionsbaum des linearen Regressionsmodells für die mittlere Goldschichtstärke der kleinen Pads im reduzierten Modellierungsdatensatz (ENIG.Reg). Die Beobachtungen werden auf Basis der Prädiktoren im Regressionsmodell in Gruppen aufgeteilt.

Dabei ergibt sich die Liegezeit im Goldbad als wichtigste Variable im Modell. Sie

spaltet den Baum in zwei Teilbäume. Im linken Teilbaum sind alle Beobachtungen mit einer Liegezeit im Goldbad von 16 Minuten und kürzer, während im rechten Teilbaum die Beobachtungen durch eine Liegezeit länger als 16 Minuten charakterisiert sind. Der rechte Teilbaum umfasst 23 Beobachtungen und wird nicht weiter aufgespalten. Besonders auffällig ist bei dem zugehörigen Boxplot der relativ hohe Median. Im Gegensatz dazu enthält der linke Teilbaum drei innere Knoten, die Maximaltemperatur im Nickelbad, die Liegezeit im Goldbad und den Nickelgehalt im Goldbad.

Zwei Kombinationen führen eher zu größeren Schichtstärken, während die anderen drei Kombinationen eher kleinere Schichtstärken zur Folge haben. Zu einer größeren Goldschichtstärke führt generell die Einstellung der Liegezeit im Goldbad über 16 Minuten oder die Kombination einer kürzeren Liegezeit im Goldbad mit einer Maximaltemperatur im Nickelbad kleiner gleich 81.4 °C. Für eine niedrigere Goldschicht sollte eine Einstellung der verbleibenden drei Kombinationen gewählt werden. Dabei ist die Maximaltemperatur im Nickelbad immer größer als 81.4 °C.

4.6.3 Modell für die mittlere Goldschichtstärke der mittleren Pads

Die nächste Tabelle 4.13 stellt die Details des Regressionsmodells für die mittlere Goldschichtstärke der mittleren Pads dar.

Tabelle 4.13: Lineares Regressionsmodell für die mittlere Goldschichtstärke der mittleren Pads im reduzierten Modellierungsdatensatz (ENIG.Reg).

$MWAuM^{-1} \sim$ SudGold_Liegezeit+AuNi+SudGold_Tmin+ChemNi_Tmax+AktivierenAu_Tmax					
	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	-2.592e-02	2.229e-02	-1.163	0.246211	
SudGold_Liegezeit	-3.387e-04	5.839e-05	-5.801	2.50e-08	***
AuNi	1.346e-06	3.247e-07	4.146	4.98e-05	***
SudGold_Tmin	-2.658e-04	6.817e-05	-3.899	0.000131	***
ChemNi_Tmax	9.033e-04	2.489e-04	3.629	0.000360	***
AktivierenAu_Tmax	-3.722e-04	1.767e-04	-2.106	0.036415	*
Observations	209				
R^2	0.303				
Adjusted R^2	0.286				
Residual Std. Error	0.001 (df = 203)				
F Statistic	17.687 *** (df = 5; 203)				
Signif. codes:	*** p<0.001 ** p<0.01 * p<0.05 . p<0.1				

Die Zielvariable wurde wieder transformiert. Im Vergleich zum Modell für die kleinen Pads in Tabelle 4.12 sind alle Variablen bis auf die Maximaltemperatur im Aktivator hoch signifikant. Die Hypothese, dass alle Regressionskoeffizienten Null sind, kann

ebenfalls verworfen werden. Das Bestimmtheitsmaß und das adjustierte Bestimmtheitsmaß sind mit 30.3% bzw. 28.6% in etwa gleich groß, aber etwas kleiner als beim Modell der kleinen Pads.

Die Residuenanalyse in Abbildung 4.35 ergibt keine besonderen Unregelmäßigkeiten. Die Residuen weisen keine Trends auf und sind annähernd normalverteilt. Keine Beobachtungen haben einen besonderen Einfluss auf die Regressionskoeffizienten und keine weitere Transformation ist nötig.

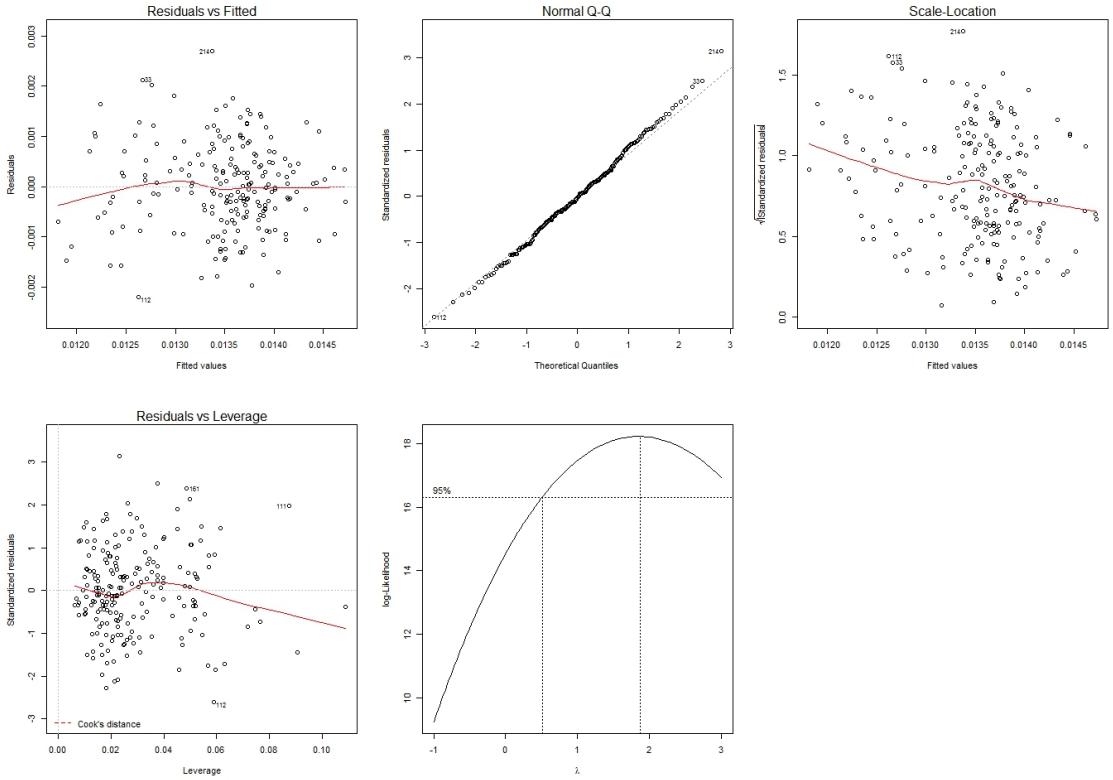

Abbildung 4.35: Modelldiagnoseplots des linearen Regressionsmodells für die mittlere Goldschichtstärke der mittleren Pads im reduzierten Modellierungsdatensatz (ENIG.Reg). Anhand verschiedener Grafiken können die Modellannahmen überprüft werden.

Wird der zugehörige Regressionsbaum des Modells in Abbildung 4.36 betrachtet, ergibt sich im Vergleich zum Regressionsbaum der mittleren Goldschichtstärke der kleinen Pads in Abbildung 4.34 ein etwas anderes Bild. Dieser Regressionsbaum ist weniger verzweigt. Die Liegezeit im Goldbad ist wiederum die wichtigste Variable und bildet die Wurzel des Baums. Allerdings existieren nur zwei weitere Knoten, die Maximaltemperatur im Nickelbad und der Nickelgehalt im Goldbad.

Wiederum sind die größten Goldschichtstärken bei einer Liegezeit länger als 16 Minuten beobachtet worden bzw. bei einer kürzeren Liegezeit in Kombination mit einer Maximaltemperatur im Nickelbad kleiner gleich 81.4 °C. Etwas kleinere Schichtstärken treten bei den zwei mittleren Blättern auf. Diese Beobachtungen haben eine Liegezeit

im Goldbad von kleiner gleich 16 Minuten und eine Maximaltemperatur im Nickelbad über 84.1 °C.

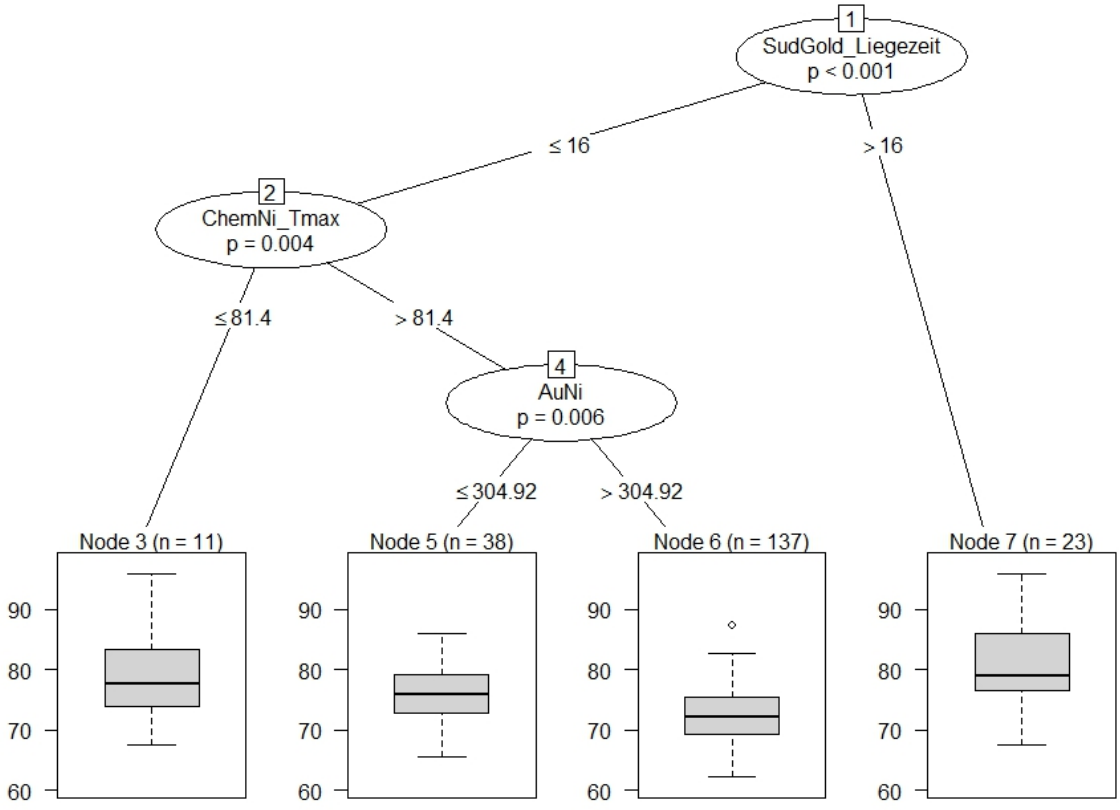

Abbildung 4.36: Regressionsbaum des linearen Regressionsmodells für die mittlere Goldschichtstärke der mittleren Pads im reduzierten Modellierungsdatensatz (ENIG.Reg). Dabei werden die Beobachtungen als verästelter Baum dargestellt.

4.6.4 Modell für die mittlere Goldschichtstärke der großen Pads

Das letzte Modell ist das Modell für die mittlere Goldschichtstärke der großen Pads. Tabelle 4.14 listet die Regressionskoeffizienten und alle wichtigen Kenngrößen auf. Dabei ist zu erkennen, dass alle Prozessparameter bis auf die Maximaltemperatur im Aktivator hoch signifikant sind. Das Bestimmtheitsmaß und das adjustierte Bestimmtheitsmaß sind mit 36.2% bzw. 34.7% ähnlich groß und etwas größer als bei den Modellen der kleinen bzw. mittleren Pads.

Die zugehörigen Grafiken zur Residuenanalyse sind in Abbildung 4.37 dargestellt.

Tabelle 4.14: Lineares Regressionsmodell für die mittlere Goldschichtstärke der großen Pads im reduzierten Modellierungsdatensatz (ENIG.Reg).

$MWAuG^{-1} \sim$ SudGold_Liegezeit+AuNi+SudGold_Tmin+ChemNi_Tmax+AktivierenAu_Tmax

	Estimate	Std. Error	t value	Pr(> |t|)	
(Intercept)	-3.692e-02	2.174e-02	-1.698	0.0910	.
SudGold_Liegezeit	-2.895e-04	5.696e-05	-5.082	8.46e-07	***
AuNi	2.172e-06	3.168e-07	6.856	8.31e-11	***
SudGold_Tmin	-3.580e-04	6.650e-05	-5.384	2.00e-07	***
ChemNi_Tmax	1.116e-03	2.428e-04	4.596	7.57e-06	***
AktivierenAu_Tmax	-3.080e-04	1.724e-04	-1.786	0.0755	.

Observations	209
R^2	0.387
Adjusted R^2	0.372
Residual Std. Error	0.001 (df = 203)
F Statistic	25.640*** (df = 5; 203)

Signif. codes: *** p<0.001 ** p<0.01 * p<0.05 . p<0.1

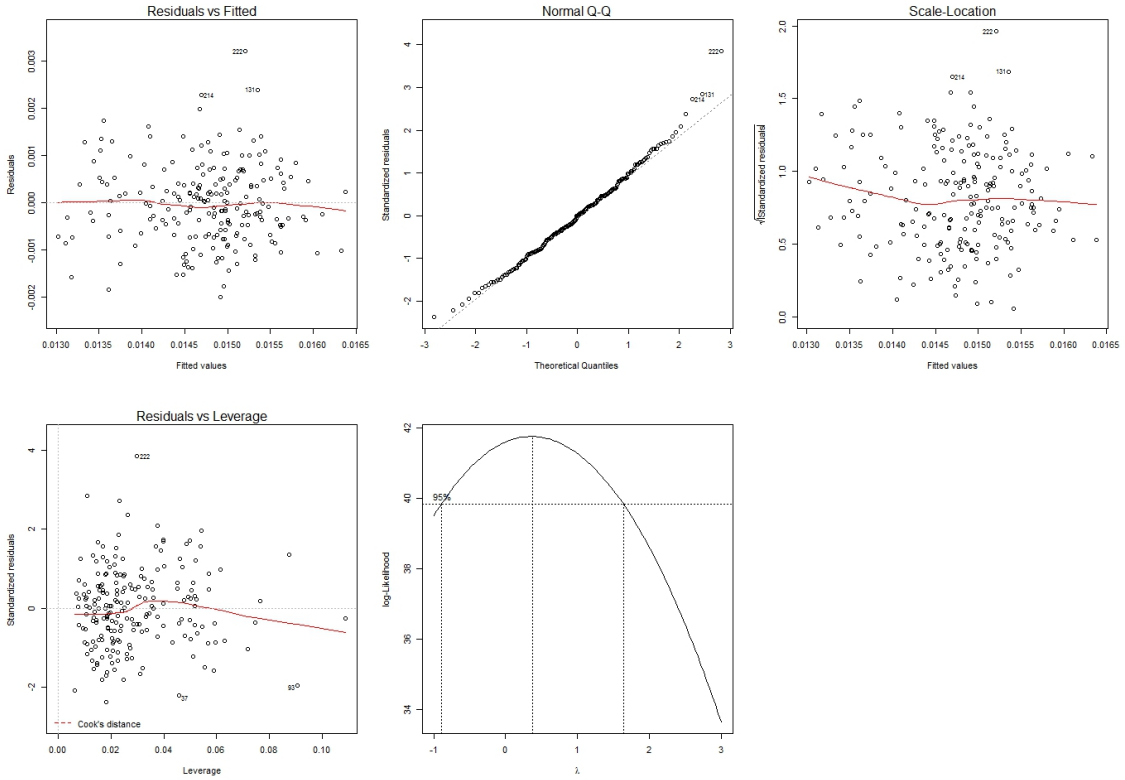

Abbildung 4.37: Modelldiagnoseplots des linearen Regressionsmodells für die mittlere Goldschichtstärke der großen Pads im reduzierten Modellierungsdatensatz (ENIG.Reg). Die Modellannahmen scheinen erfüllt zu sein.

Die Residuen zeigen keinen systematischen Trend und die standardisierten Residuen scheinen normalverteilt zu sein. Ebenso ist die Varianzhomogenität der Residuen gegeben und die Regressionskoeffizienten hängen nicht von einzelnen Beobachtungen ab. Da die Zielvariable schon transformiert wurde, zeigt auch der Boxcox-Test keine Besonderheiten.

Abbildung 4.38 stellt den zugehörigen Regressionsbaum dar. Die Liegezeit im Goldbad stellt die Wurzel dar und ist somit die wichtigste Variable. Sie spaltet den Baum in zwei Teilbäume.

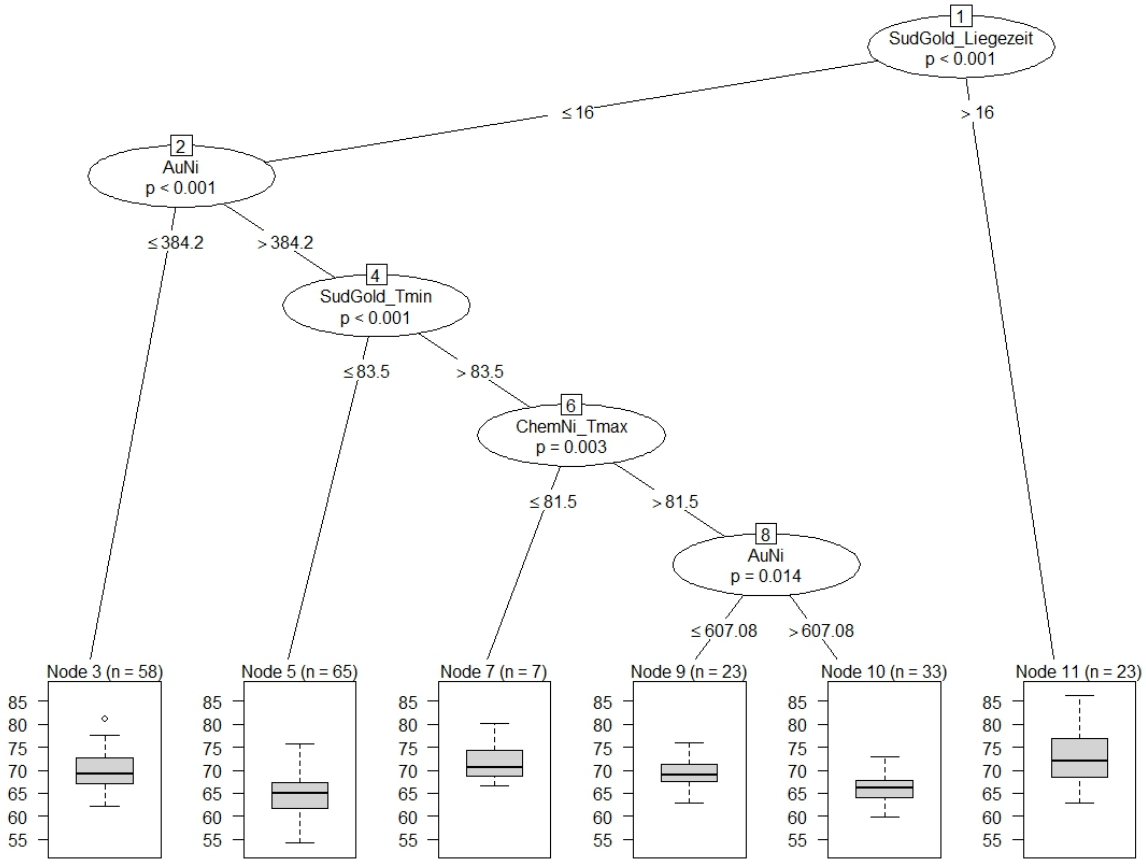

Abbildung 4.38: Regressionsbaum des linearen Regressionsmodells für die mittlere Goldschichtstärke der großen Pads im reduzierten Modellierungsdatensatz (ENIG.Reg).

Der rechte Teilbaum wird nicht aufgespalten und beinhaltet 23 Beobachtungen, die durch eine Liegezeit über 16 Minuten charakterisiert sind. Die mittleren Goldschichtstärken sind dabei eher größer. Im Gegensatz dazu enthält der linke Teilbaum alle restlichen Beobachtungen und wird mittels vier innerer Knoten weiter aufgespalten. Die inneren Knoten bestehen aus dem Nickelgehalt im Goldbad, der Minimaltemperatur im Goldbad und der Maximaltemperatur im Nickelbad. Die Verteilung der mittleren Schichtstärken ist je nach Kombination der Parametereinstellungen unterschiedlich. Die kleinsten Schichtstärken resultieren am ehesten aus der Kombination

einer kürzeren Liegezeit, einem größeren Nickelgehalt im Goldbad und einer niedrigeren Temperatur im Goldbad. Falls die Temperatur im Goldbad höher ist, dann sollte für eine dünnere Goldschicht auch der Nickelgehalt im Goldbad größer sein.

Abschließend sind in Abbildung 4.39 noch einmal die drei Regressionsbäume zusammengefasst. Dabei sind jene Blätter mit kleineren Goldschichtstärken türkis markiert und Blätter mit größeren Goldschichtstärken pink.

Hierbei zeigt sich, dass die Liegezeit im Goldbad jeweils die Wurzel der drei Regressionsbäume bildet. Die rechten Teilbäume, die eine längere Liegezeit im Goldbad charakterisieren, werden nicht weiter aufgespalten und führen zu größeren Goldschichtstärken. Die linken Teilbäume beinhalten unterschiedlich viele innere Knoten. Der Regressionsbaum für die mittleren Pads weist zwei innere Knoten auf, während die Regressionsbäume der kleinen bzw. großen Pads drei bzw. vier innere Knoten besitzen. Die Prozessparameter Maximaltemperatur im Nickelbad und Nickelgehalt im Goldbad kommen dabei in allen drei Bäumen vor. Unterdessen ist ein innerer Knoten im Modell der kleinen Pads die Liegezeit im Goldbad, wie die Wurzel, und ein innerer Knoten im Modell der großen Pads die Minimaltemperatur im Goldbad.

Als ein Beispiel resultiert die Einstellungskombination der Prozessparameter

(Liegezeit im Goldbad $\leq$ 16 Minuten) und
(Nickelgehalt im Goldbad > 607.08 mg/l) und
(Minimaltemperatur im Goldbad > 83.5 $^{\circ}C$) und
(Maximaltemperatur im Nickelbad > 81.5 $^{\circ}C$)

in allen drei Regressionsbäumen in kleineren mittleren Goldschichtstärken. Wird eine größere Goldschichtstärke gefordert, führt zum Beispiel eine längere Liegezeit im Goldbad über 16 Minuten bei allen Pad-Größen zum Ziel.

Zusammenfassend ist festzustellen, dass zur gemeinsamen Modellierung der drei Zielvariablen, die zur Vorhersage verwendet werden soll, nur fünf Parameter geeignet ist. Die Prädiktoren sind:

- die Liegezeit im Goldbad,
- der Nickelgehalt im Goldbad,
- die Minimaltemperatur im Goldbad,
- die Maximaltemperatur im Nickelbad und
- die Maximaltemperatur im Aktivator.

Zur besseren Anpassung des Modells wurden die Zielvariablen transformiert. Der Erklärungsgehalt der Regressionsmodelle ist nicht stark ausgeprägt. Der Grund der geringen adjustierten Bestimmtheitsmaße der Regressionsmodelle ist durch die explorative Analyse deutlich erkennbar. Die Goldschichtstärke weist große Schwankungen bei festen Parametereinstellungen auf. Damit kann die Punktschätzung des Regressionsmodells keine besseren Prognosen liefern. Auch sind die Standard Residual Errors und Messfehler der Goldschichtstärken ähnlich groß, vgl. Riedler, 2015. Dadurch liegt die Genauigkeit der Modelle fast im Messfehlerbereich. Jedoch liefern die Regressionsbäume je nach Parametereinstellungen unterschiedliche Verteilungen, die für eine Optimierung der Goldschichtstärke herangezogen werden können.

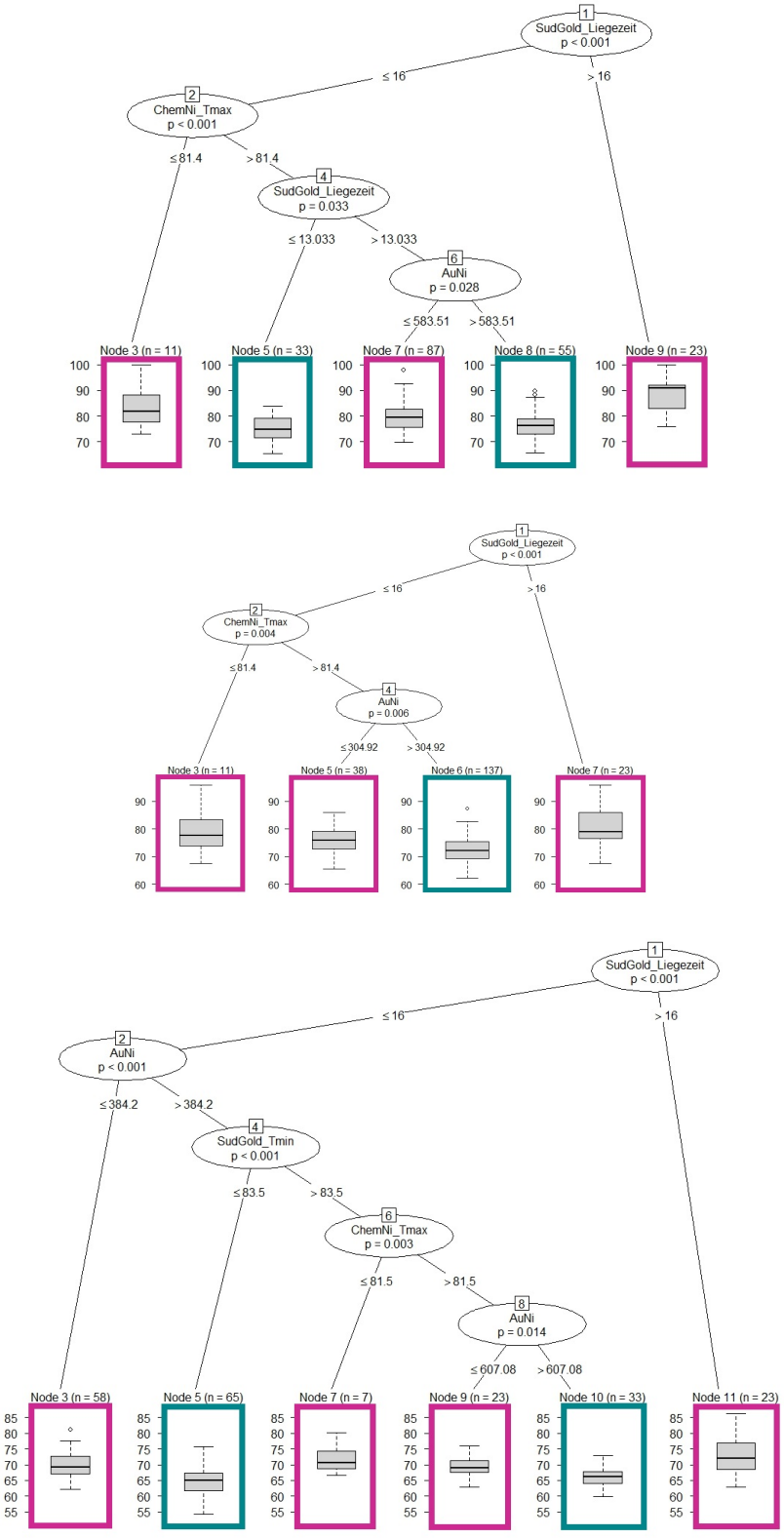

Abbildung 4.39: Regressionsbäume der drei Zielvariablen: mittlere Goldschichtstärke der kleinen/mittleren/großen Pads (ENIG.Reg). Blätter, deren Parametereinstellungen zu kleineren Goldschichtstärken führen, sind türkis markiert. Blätter, deren Parametereinstellungen zu größeren Goldschichtstärken führen, sind pink markiert.

4.7 Validierung der Regressionsmodelle

Zur Validierung des multivariaten Modells wurden weitere Messungen durchgeführt. Der Schnitt zwischen den Modellierungsdaten zur Modellierung der Regressionsmodelle (ENIG.Reg) und den Validierungsdaten zur Modellvalidierung (ENIG.Vali) wurde zufällig gesetzt.

Im Zeitraum vom 01.03.2015 bis 07.04.2015 wurden 132 Beobachtungen zum Zweck der Validierung aufgezeichnet. Nachdem der Datensatz bereinigt wurde, blieben 70 vollständige Beobachtungen für die Beurteilung der Modellgüte übrig, vgl. Abbildung 4.40. Die meisten nicht verwendbaren Beobachtungen mussten in Folge inkonsistenter Datensätze verworfen werden. Weitere 10 Beobachtungen konnten aufgrund fehlender Plausibilität der Messungen der Schichtstärken nicht weiter behandelt werden.

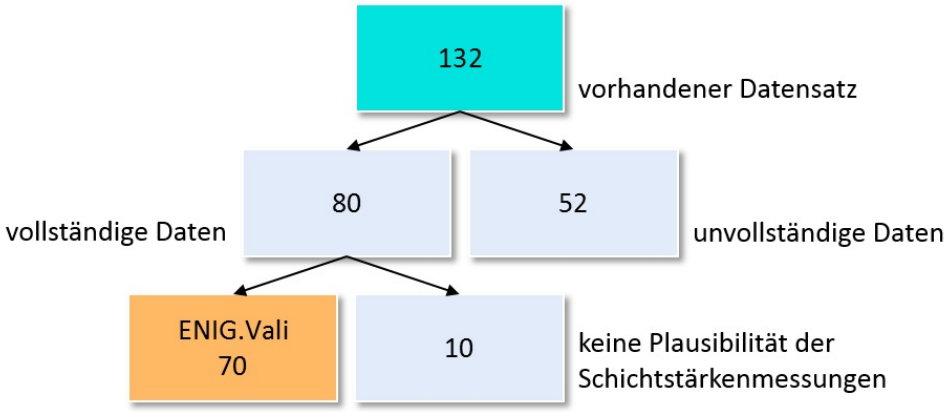

Abbildung 4.40: Validierungsdatenbasis. Der Datensatz (ENIG.Vali) wurde für die Validierung der Regressionsmodelle herangezogen.

4.7.1 Multivariates Modell für die mittlere Goldschichtstärke

Wie sich das entwickelte Regressionsmodell auf neue Daten anwenden lässt, zeigt Abbildung 4.41. Darin sind die Beobachtungen der mittleren Goldschichtstärke der unterschiedlichen Pad-Größen des Validierungsdatensatzes als Punkte eingetragen. Anhand des multivariaten Regressionsmodells, das auf Basis der Modellierungsdaten entstanden ist, können Vorhersagen berechnet werden. Diese sind in pink eingezeichnet. Ebenso sind die zugehörigen 95%-Vorhersageintervalle illustriert. Mit Hilfe des Abstands zwischen den beobachteten und den vorhergesagten Werten (Residuen) kann das Modell beurteilt werden.

Die Differenz der beobachteten und vorhergesagten Werte des Validierungsdatensatzes liegt in den Intervallen:

- $[-26.58\ \mu m; 22.77\ \mu m]$ beim Modell der kleinen Pads ($\hat{\sigma} = 9.06$),
- $[-24.09\ \mu m; 20.55\ \mu m]$ beim Modell der mittleren Pads ($\hat{\sigma} = 8.15$) und
- $[-14.11\ \mu m; 24.37\ \mu m]$ beim Modell der großen Pads ($\hat{\sigma} = 7.95$).

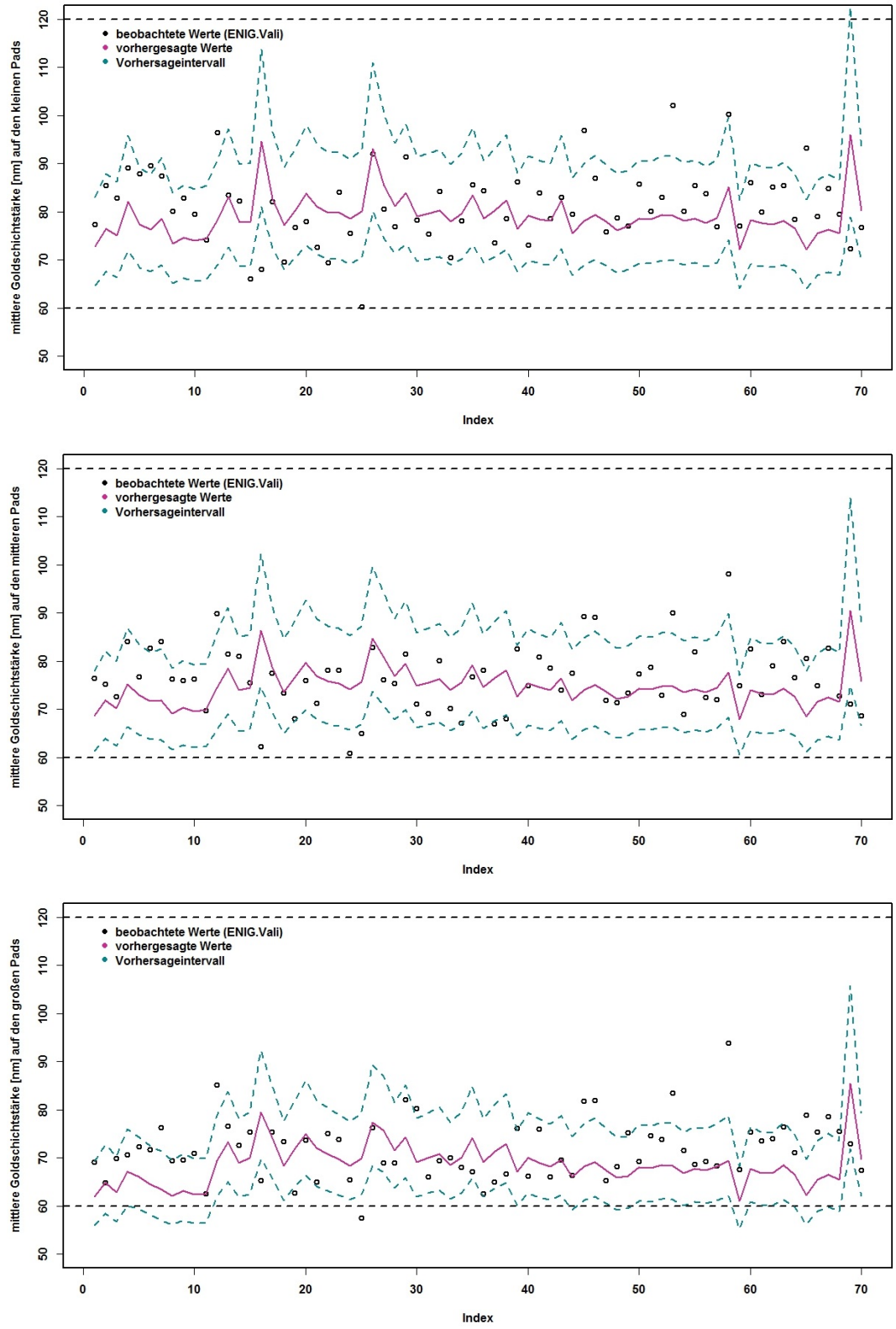

Abbildung 4.41: Validierung des multivariaten Regressionsmodells getrennt nach Pad-Größe (ENIG.Reg, ENIG.Vali). Der absolute Abstand zwischen den beobachteten Werten des Validierungsdatensatzes und den vorhergesagten Werten liegt im Intervall $[0\ \mu m; 26.58\ \mu m]$.

4.7.2 Modifiziertes multivariates Modell für die mittlere Goldschichtstärke

Der Datensatz ENIG.gesamt ist die Kombination der Datensätze ENIG.Reg und ENIG.Vali. Er kombiniert die bereinigten Daten des neuen Datensatzes ohne Ausreißer (ENIG.Reg) mit den bereinigten Daten für die Validierung des Modells (ENIG.Vali). Wird der gesamte Datensatz zur Modellierung verwendet, ändern sich die signifikanten Parameter. Die Maximaltemperatur im Nickelbad als auch im Aktivator leisten keinen signifikanten Beitrag mehr zur Beschreibung der mittleren Goldschichtstärke. Die Parameter Liegezeit im Goldbad, Nickelgehalt im Goldbad und Temperatur im Goldbad sind stabil. Aus diesem Grund wurden die nicht-stabilen Prozessparameter herausgenommen und nur die drei stabilen Parameter zur Modellierung herangezogen.

Tabelle 4.15 gibt einen Überblick über das reduzierte Modell, basierend auf dem Datensatz ENIG.Reg.

Tabelle 4.15: Reduziertes multivariates Regressionsmodell für die mittlere Goldschichtstärke im reduzierten Modellierungsdatensatz (ENIG.Reg).

$MWAuK^{-1}$, $MWAuM^{-1}$, $MWAuG^{-1} \sim$ SudGold_Liegezeit+AuNi+SudGold_Tmin			
Coefficients:			
	$MWAuK^{-1}$	$MWAuM^{-1}$	$MWAuG^{-1}$
(Intercept)	3.740e-02	4.054e-02	4.920e-02
SudGold_Liegezeit	-4.795e-04	-3.616e-04	-3.133e-04
AuNi	1.036e-06	1.251e-06	2.126e-06
SudGold_Tmin	-2.247e-04	-2.733e-04	-3.752e-04

Die nächsten drei Grafiken in Abbildung 4.42 stellen die Modellvalidierung des reduzierten multivariaten Regressionsmodells grafisch dar. Dazu sind die beobachteten Werte des Validierungsdatensatzes ENIG.Vali als Punkte aufgetragen. Die vorhergesagten Werte, die auf Basis des Modells aus Tabelle 4.15 berechnet wurden, sind in pink dargestellt und das zugehörige 95% Vorhersageintervall in türkis.

Die Differenz der beobachteten und vorhergesagten Werte des Validierungsdatensatzes liegt in den Intervallen:

- $[-24.04\ \mu m; 24.90\ \mu m]$ beim Modell der kleinen Pads ($\hat{\sigma} = 8.24$),
- $[-22.52\ \mu m; 19.38\ \mu m]$ beim Modell der mittleren Pads ($\hat{\sigma} = 7.19$) und
- $[-13.35\ \mu m; 23.10\ \mu m]$ beim Modell der großen Pads ($\hat{\sigma} = 7.26$).

Die Intervalle sind im Vergleich zum vorangegangenen Modell etwas kleiner. Dies zeigt deutlich, dass die nicht-stabilen Prozessparameter die Prognose verschlechtern.

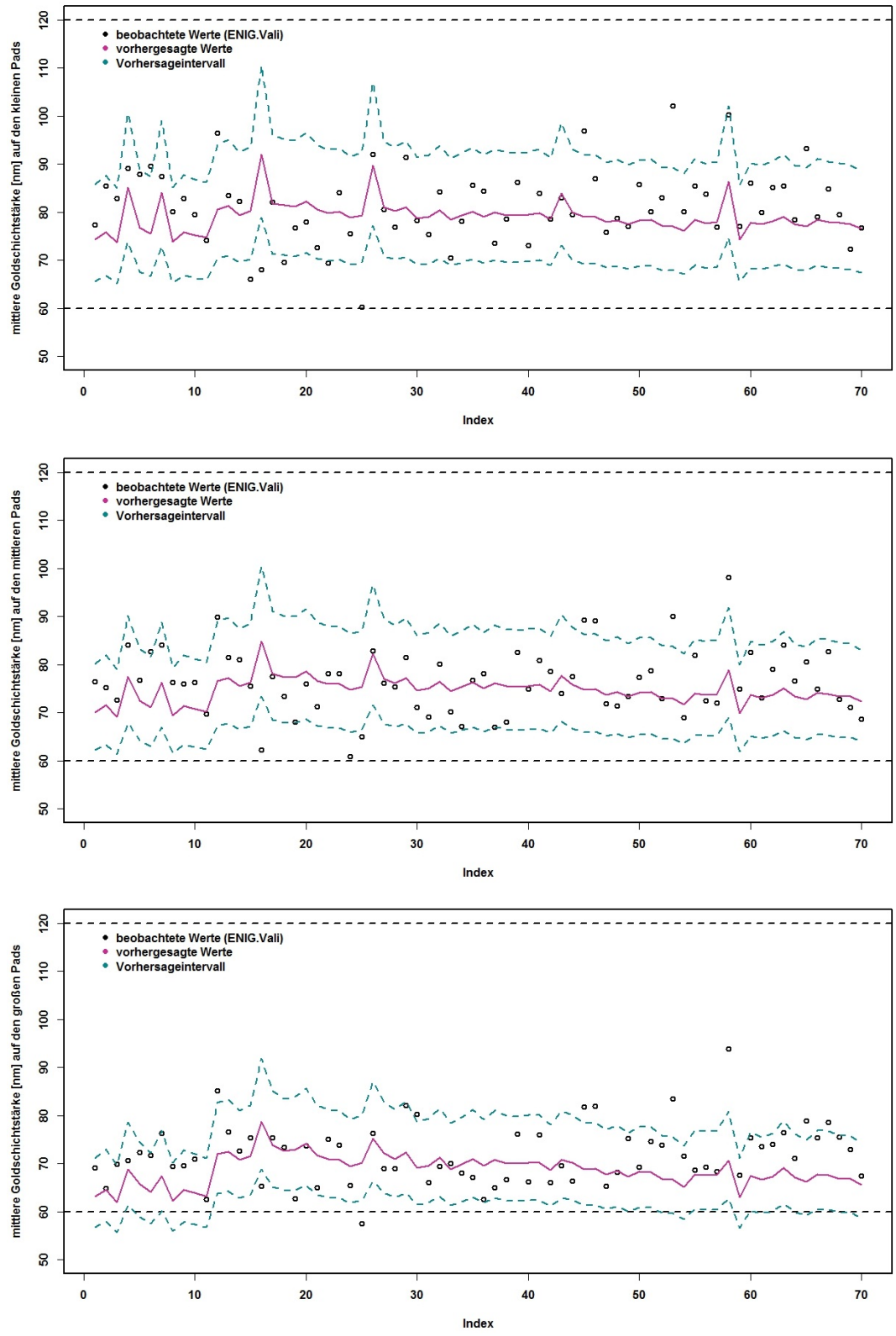

Abbildung 4.42: Validierung des modifizierten multivariaten Regressionsmodells getrennt nach Pad-Größe (ENIG.Reg,ENIG.Vali). Der absolute Abstand zwischen den beobachteten Werten des Validierungsdatensatzes und den vorhergesagten Werten liegt im Intervall [0 μm; 24.90 μm].

Abschließend liefert Abbildung 4.43 die zugehörigen Regressionsbäume zum reduzierten multivariaten Regressionsmodell aus Tabelle 4.15. Die Blätter, deren Parametereinstellungen eine niedrigere Goldschicht zur Folge haben, sind türkis markiert, während pink markierte Blätter eher höhere Goldschichten aufweisen.

Wird eine Minimierungsstrategie verfolgt, stellt die Einstellungskombination der Prozessparameter

(Liegezeit im Goldbad $\leq$ 16 Minuten) und
(Nickelgehalt im Goldbad > 583.51 mg/l) und
(Minimaltemperatur im Goldbad $\leq$ 83.5 °C)

eine Möglichkeit dar, um kleinere mittlere Goldschichtstärken in allen drei Regressionsbäumen zu bekommen. Soll die Goldschichtstärke größer sein, führt zum Beispiel eine längere Liegezeit im Goldbad über 16 Minuten bei allen Pad-Größen zum Ziel.

Zusammengefasst eignen sich die Prozessparameter zur Beschreibung einzelner Datensätze, können aber nur unter Vorbehalt für eine Vorhersage neuer Beobachtungen verwendet werden. Dies zeigt die Validierung des Regressionsmodells des Datensatzes ENIG.Reg. Demnach sind die Parameter, die zur Modellierung eines spezifischen Datensatzes passend erscheinen, nicht zur Modellierung anderer Beobachtungen übertragbar.

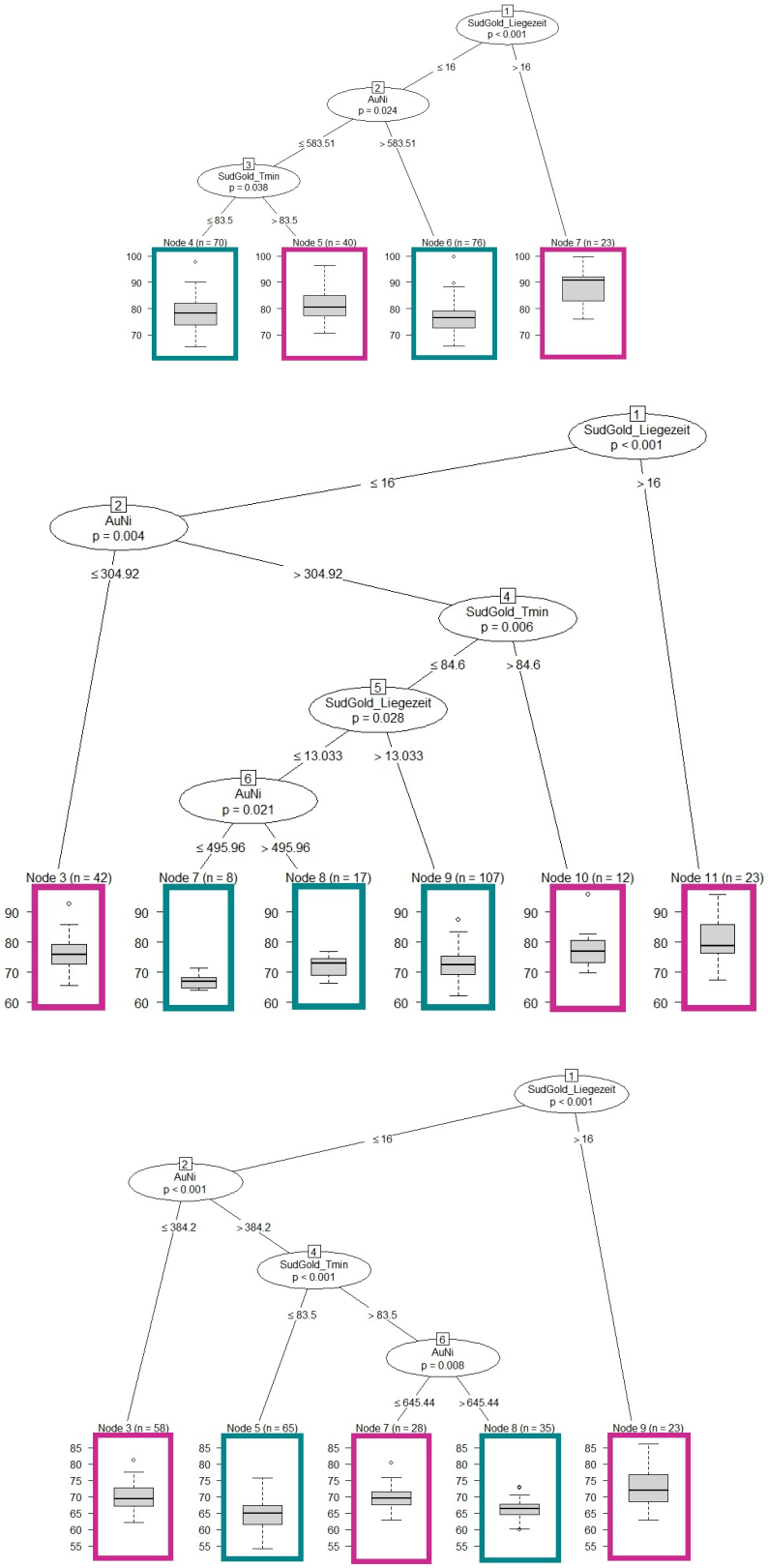

Abbildung 4.43: Reduzierte Regressionsbäume der drei Zielvariablen: mittlere Goldschichtstärke der kleinen/mittleren/großen Pads (ENIG.Reg). Blätter, deren Parametereinstellungen zu kleineren Goldschichtstärken führen, sind türkis markiert. Blätter, deren Parametereinstellungen zu größeren Goldschichtstärken führen, sind pink markiert.

5 Zusammenfassung

Ziel dieser Arbeit ist die Untersuchung der Einflüsse von ausgewählten Prozessparametern auf die Goldschichtstärke im ENIG-Prozess. Die statistische Analyse basiert auf explorativen Analysen und multivariaten multiplen Regressionsmodellen. Die Ergebnisse sollen vorhandenes Prozesswissen erweitern und zukünftige Prozessentscheidungen erleichtern. Die Aufgabenstellung umfasst die Fragen:

- Welche Prozessparameter haben einen Einfluss auf die Nickel- und Goldschichtstärke?
- Wie sollen jene Parameter eingestellt werden um den Edelmetallverbrauch zu minimieren?

Die statistischen Analysen ergaben folgendes Bild:

Die Qualität der Daten spielt eine große Rolle in statistischen Auswertungen, denn diese bilden die Grundlage. Ist die Qualität der Daten nicht entsprechend gut, kann eine statistische Analyse nur bedingt korrekte Auswertungen liefern. Falls verfälschte Daten enthalten sind, können Fehlinterpretationen die Folge sein.

Die untersuchten unbearbeiteten Datensätze waren nur bedingt nutzbar. Durch Bereinigung von fehlenden Werten, Schreibfehlern, falschen Referenzen, unterschiedlichen Labels, widersprüchlichen Werten, unterschiedlichen Einheiten, unterschiedlichen Genauigkeiten, Duplikaten, strukturellen Inhomogenitäten, ..., wurden die Datensätze aufbereitet. Enthaltene Messfehler konnten im Nachhinein nicht entfernt werden.

Die explorative Analyse der Daten ergibt ein homogenes Bild der Schichtstärken. Diese sind bis auf Ausnahmen im vorgegebenen Spezifikationsbereich, sogar eher im Idealbereich. Die bivariate Analyse ergibt, dass die Zusammenhänge zwischen der Goldschichtstärke und den Prozessparametern nicht sehr stark ausgeprägt sind. Schwankungen der Goldschichtstärke sind bei festen Parametereinstellungen die Regel. Die Prozessparameter mit dem stärksten Zusammenhang zur mittleren Goldschichtstärke sind die Liegezeit im Goldbad und die Temperaturen im Goldbad. Die Korrelationen zwischen den einzelnen Prozessparametern sind ebenfalls nicht auffallend. Auch bei einer mehrdimensionalen Betrachtung sind kaum Strukturen und Muster erkennbar.

Regressionsmodelle können verschiedene Ziele haben. Einerseits als erklärendes Modell zur Beschreibung von Datensätzen und zum Verständnis der Beziehung der Variablen oder andererseits zur Vorhersage neuer Daten (Faraway, 2004). Die vorgeschlagenen Regressionsmodelle sind zur Beschreibung der Datensätze passabel. Die Modellstreuung ist nur geringfügig größer als der Messfehler (Riedler, 2015). Zur Vorhersage neuer Daten ist das präsentierte Regressionsmodell für den Modellierungsdatensatz nur beschränkt geeignet. Die Validierung ergab einige nicht stabile Parameter,

die anschließend entfernt wurden. Trotz der geringen Zusammenhänge der mittleren Goldschichtstärke mit den Prozessparametern und der schwachen Modellgüte, liefern die vorgestellten Regressionsbäume einen Vorschlag zu Parametereinstellungen für eine minimale/maximale Goldschichtstärke.

Der Grund der nicht zufriedenstellenden Genauigkeit der Vorhersagen und damit des Regressionsmodells ist auf der einen Seite durch die explorative Analyse klar ersichtlich, da die Goldschichtstärke große Schwankungen bei festen Parametereinstellungen aufweist. Auf der anderen Seite ist die Ungenauigkeit der Vorhersagen noch nicht geklärt. Entweder werden einflussreiche Prozessparameter nicht gemessen, die Analysen von Messfehlern der Prozessparameter und Schichtstärken stark beeinflusst und verzerrt oder die Stärke der Goldschicht basiert auf einem zufälligen Prinzip.

Für weitere Analysen wird ein Versuchsplan empfohlen. Damit können Lücken im Datensatz aufgebessert werden, um Parametereinstellungen, die in der realen Produktion zu selten vorkommen, zu überprüfen. Ebenfalls können damit die Einstellungsgrenzen der Prozessparameter genau analysiert werden.

Fragen, die sich im Laufe der Analysen stellten und als Basis weiterführender Analysen dienen können, sind:

- Sollte die Messprozedur der Prozessparameter geändert werden bzw. werden an der richtigen Stelle zum richtigen Zeitpunkt die richtigen Parameter gemessen?
- Gibt es nicht beobachtete einflussreiche Prozessparameter?
- Was ist die Ursache für die großen Schwankungen der mittleren Goldschichtstärke bei gleichbleibenden Parametereinstellungen?

Zusätzlich sollte die Datenqualität durch eine adaptierte Messprozedur, konsistente Datensätze, Verminderung von Messfehlern, etc. verbessert werden.

Anhang

Statistische Kenngrößen

Sämtliche Definitionen sind in den Vorlesungsmitschriften von Stadlober, 2006, und Stadlober, 2008, bzw. dem Skript von Stadlober, 2013, zu finden.

Es seien $x_1, \ldots, x_n$ Realisationen der Stichprobe $X_1, \ldots, X_n$ und $x_{(1)} \leq \cdots \leq x_{(n)}$ Realisationen der geordneten Stichprobe $X_{(1)} \leq \cdots \leq X_{(n)}$.

Definition 1 *Arithmetisches Mittel | Das **arithmetische Mittel** einer Stichprobe ist definiert als:*

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Es ist eine Kennzahl zur Beschreibung der Lage.

Definition 2 *Quantil | Das **p-te Quantil** einer Stichprobe ist definiert als:*

$$q_p = (1-g)x_{(\lfloor (n-1)p \rfloor + 1)} + g x_{(\lfloor (n-1)p \rfloor + 2)},$$

$$g = (n-1)p - \lfloor (n-1)p \rfloor.$$

*Das p-te Quantil hat die Eigenschaft, dass mindestens ein relativer Anteil von p Daten kleiner gleich q_p ist und höchstens ein relativer Anteil (1-p) größer als q_p. Das 1. Quartil, der **Median** und 3. Quartil sind besondere Quantile und entsprechen $q_{0.25}$, $q_{0.50}$ und $q_{0.75}$.*

Definition 3 *Interquartilsabstand | Der **Interquartilsabstand** ist der Abstand zwischen dem 1. und dem 3. Quartil einer Stichprobe:*

$$IQR = q_{0.75} - q_{0.25}.$$

Je größer der Interquartilsabstand ist, desto größer ist die Streuung der mittleren 50% der Daten.

Definition 4 *Spannweite | Die **Spannweite** einer Stichrobe ist die Differenz des Maximums der Stichprobe $x_{(n)}$ und des Minimums der Stichprobe $x_{(1)}$:*

$$spw = x_{(n)} - x_{(1)}.$$

.

Definition 5 *Standardabweichung, Varianz | Die **Standardabweichung** einer Stichprobe ist definiert als:.*

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Sie gibt die durchschnittliche Abweichung vom Mittelwert an.

*Die **Varianz** einer Stichprobe ergibt sich als Quadrat der Standardabweichung selbiger:*

$$Var(x) = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Definition 6 *Kovarianz | Die **Kovarianz** zweier Stichproben X und Y ist definiert als:*

$$Cov\,(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Sie ist ein Maß für die gemeinsame Variabilität zweier Merkmale.

Definition 7 *Korrelationskoeffizient | Der **Korrelationskoeffizient** der Stichprobe X und Y ist definiert als*

$$Corr(x,y) = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

X und Y sind

- *positiv korrelliert, falls* $r_{xy} > 0$,
- *unkorrelliert, falls* $r_{xy} = 0$,
- *negativ korrelliert, falls* $r_{xy} < 0$.

Literatur

Akkerboom, H. (2012). *Wirtschaftsstatistik im Bachelor: Grundlagen und Datenanalyse*. 3. Auflage. Gabler Verlag, Wiesbaden (siehe S. 7).

Atotech GmbH (2014). *Präsentation: Atotech´s ENIG Processes. Mechanism* (siehe S. 4, 5).

AT&S AG (2015a). *AT&S Onlineführung*. Zugriff: Mai 2015. URL: http://www.ats.net/de/unternehmen/erlebniswelt/online-fuhrung/ (siehe S. 3, 4).

AT&S AG (2015b). *Die Zukunft stellt viele große Fragen. Geschäftsbericht 2014/2015*. AT&S Austria Technologie & Systemtechnik Aktiengesellschaft (siehe S. 3).

AT&S AG (2015c). *Unternehmensinformation*. Zugriff: April 2015. URL: http://www.ats.net/de (siehe S. 3).

Bankhofer, U. und Vogel, J. (2008). *Datenanalyse und Statistik: Eine Einführung für Ökonomen im Bachelor*. Gabler Verlag, Wiesbaden (siehe S. 25).

Breitwieser, D. (2015). *Interner Report*. AT&S Austria Technologie & Systemtechnik Aktiengesellschaft (siehe S. 30, 51, 52).

Burkschat, M., Cramer, E. und Kamps, U. (2012). *Beschreibende Statistik: Grundlegende Methoden der Datenanalyse*. 2. Auflage. Springer Verlag, Berlin Heidelberg (siehe S. 7, 12, 21).

Cleff, T. (2011). *Deskriptive Statistik und moderne Datenanalyse: Eine computergestützte Einführung mit Excel, PASW (SPSS) und STATA*. 2. Auflage. Gabler Verlag, Wiesbaden (siehe S. 7).

Eklund, A. (2015). *The Bee Swarm Plot, an Alternative to Stripchart*. Zugriff: Juni 2015. URL: http://cran.r-project.org/web/packages/beeswarm/beeswarm.pdf (siehe S. 9).

Fahrmeir, L., Kneib, T. und Lang, S. (2009). *Regression: Modelle, Methoden und Anwendungen*. 2. Auflage. Springer Verlag, Berlin Heidelberg (siehe S. 7, 13, 16–19, 21, 23, 24).

Faraway, J. (2004). *Practical Regression and ANOVA using R*. Zugriff: April 2015. URL: https://cran.r-project.org/doc/contrib/Faraway-PRA.pdf (siehe S. 12, 14, 17, 40, 45, 85).

Fox, J. und Weisberg, S. (2011). *Multivariate Linear Models in R*. 2nd Edition. Sage Publications, Thousand Oaks, CA (siehe S. 26).

Friedl, H. (2008). *Regressionsanalyse*. Vorlesungsmitschrift. Institut für Statistik, Technische Universität Graz (siehe S. 7, 13, 16, 21, 22, 24).

Friedl, H. (2014). *Generalisierte lineare Modelle*. Vorlesungsmitschrift. Institut für Statistik, Technische Universität Graz (siehe S. 20).

Grömping, U. (2009). »Variable Importance Assessment in Regression: Linear Regression versus Random Forest«. In: *The American Statistician* 63.4, S. 308–319 (siehe S. 25).

Gross, J. (2010). *Grundlegende Statistik mit R: Eine anwendungsorientierte Einführung in die Verwendung der Statistik Software R*. Vieweg+Teubner Verlag, Wiesbaden (siehe S. 22).

Hothorn, T. u. a. (2015). *A Laboratory for Recursive Partytioning*. Zugriff: Juni 2015. url: `http://cran.r-project.org/web/packages/party/party.pdf` (siehe S. 25, 47).

Kohn, W. (2005). *Statistik: Datenanalyse und Wahrscheinlichkeitsrechnung*. Springer Verlag, Berlin Heidelberg (siehe S. 12–14).

Kohn, W. und Öztürk, R. (2013). *Statistik für Ökonomen: Datenanalyse mit R und SPSS*. 2. Auflage. Springer Verlag, Berlin Heidelberg (siehe S. 7, 8, 12).

Kronthaler, F. (2014). *Statistik angewandt: Datenanalyse ist (k)eine Kunst*. Springer Verlag, Berlin Heidelberg (siehe S. 12).

Lumley, T. (2015). *Regression Subset Selection*. Zugriff: Juni 2015. url: `http://cran.r-project.org/web/packages/leaps/leaps.pdf` (siehe S. 23, 40, 67).

Maitra, R. (2013). *Multivariate Multiple Regression*. Zugriff: Juni 2015. url: `http://www.public.iastate.edu/~maitra/stat501/lectures/MultivariateRegression.pdf` (siehe S. 26).

McCullagh, P. und Nelder, J. (1989). *Generalized linear models*. 2nd Edition. Chapman und Hall, London (siehe S. 12).

Naumann, F. (2007). »Datenqualität«. In: *Informatik-Spektrum* 30.1, S. 27–31 (siehe S. 32).

Pruscha, H. (2006). *Statistisches Methodenbuch: Verfahren, Fallstudien, Programmcodes*. Springer Verlag, Berlin Heidelberg (siehe S. 11, 18).

Riedler, M. (2015). *Interner Report*. AT&S Austria Technologie & Systemtechnik Aktiengesellschaft (siehe S. 55, 77, 85).

Ripley, B. u. a. (2015). *Support Functions and Datasets for Venables and Ripley's MASS*. Zugriff: Juni 2015. URL: http://cran.r-project.org/web/packages/MASS/MASS.pdf (siehe S. 20).

Schendera, C. (2007). *Datenqualität mit SPSS*. Oldenbourg Verlag, München (siehe S. 32, 33, 54).

Stadlober, E. (2006). *Statistik*. Vorlesungsmitschrift. Institut für Statistik, Technische Universität Graz (siehe S. 8, 9, 11, 89).

Stadlober, E. (2008). *Angewandte Statistik*. Vorlesungsmitschrift. Institut für Statistik, Technische Universität Graz (siehe S. 7, 10, 67, 89).

Stadlober, E. (2013). *Angewandte Statistik*. Vorlesungsskript. Institut für Statistik, Technische Universität Graz (siehe S. 89).

Toutenburg, H. u. a. (2009). *Arbeitsbuch zur deskriptiven und induktiven Statistik*. 2. Auflage. Springer Verlag, Berlin Heidelberg (siehe S. 10).

Wei, T. (2015). *Visualization of a correlation matrix*. Zugriff: Juni 2015. URL: http://cran.r-project.org/web/packages/corrplot/corrplot.pdf (siehe S. 38).

Wermuth, N. und Streit, R. (2007). *Einführung in statistische Analysen: Fragen beantworten mit Hilfe von Daten*. Springer Verlag, Berlin Heidelberg (siehe S. 7, 11).