Prinzipien der Marktdatenvalidierung

Dr. André Miemiec[1] und Kerstin Steinberg

FRAME Consulting GmbH

Gabriel-Max-Straße 12

10245 Berlin

Abstract

In diesem Artikel wird ein Marktdatenvalidierungsprozeß vom Datenbezug bis hin zur Erstellung einer qualitätsgesicherten goldenen Kopie skizziert. Das Herzstück bildet dabei der Zeitreihenbereinigungsprozeß.
Die Relevanz eines adäquaten Zeitreihenbereinigungsprozesses ergibt sich aus seiner zentralen Bedeutung für mindestens zwei End-of-Day-Prozesse: Für den Mark-to-Market-Lauf zur Produktion von Fair Values und für den Risikolauf zur Produktion von aussagekräftigen Risikokennzahlen.
Dieser Artikel wird sich in der Darstellung auf die Zeitreihenbereinigung für die Zwecke des Risikolaufs konzentrieren.
Der hier dargestellte Prozeß beinhaltet neben univariaten auch multivariate Ausreißertests. Die statistischen Ausreißertests lassen sich i.d.R. als Hypothesentests auffassen. Eine geeignete Hintereinanderschaltung von univariaten und multivariaten Tests in Form einer Probe aufs Exempel macht es möglich, Fehler 1. und 2. Art bei den univariaten Tests getrennt zu identifizieren.
Dieses Verfahren bietet den Mehrwert, sowohl Ausreißer, die das Volatilitätsniveau verzerren, als auch Ausreißer, die die Korrelationsstruktur verzerren, zu identifizieren. Beide Effekte können sich nachdrücklich auf die von einem Risikomodell produzierten Risikokennzahlen auswirken.

Keywords: Market data validation, outliers, imputation

[1] Corresponding author: andre.miemiec@frame-consult.de

1 Einleitung

Qualitätsgesicherte Marktdaten sind eine der Grundvoraussetzungen für die Durchführung des End-of-Day-Prozesses (EOD) einer Bank. Sie sind nötig, um sowohl aussagekräftige Mark-to-Market-Werte (MtM) als auch Risikokennzahlen ableiten zu können. Für die Erstellung von Risikoprognosen ist zusätzlich die Bestimmung der Risikoverteilung erforderlich. Diese kann entweder durch Parameterschätzung einer vorgegebenen Verteilung oder die Verwendung der historischen Verteilung der Risikofaktoren dargestellt werden. Für die Konstruktion beider Darstellungen von Risikoverteilungen ist das Vorliegen qualitätsgesicherter Marktdatenzeitreihen erforderlich.

Marktdaten werden von vielen Anbietern zur Verfügung gestellt. Die Daten aller Anbieter können fehlerbehaftet sein; sei es, daß falsche Werte, sei es, daß keine Werte für bestimmte Datenpunkte angeliefert werden. Die Verwendung von fehlerhaften Marktdaten in Risikomodellen kann im Extremfall zu falschen Einschätzungen der Risikodiversifikation bzw. Risikokonzentration führen. Eine besondere Rolle wird in diesem Zusammenhang die Korrelation von Risikofaktoren spielen.
Der ökonomische Mehrwert eines konsistenten Risikomodells besteht u.a. darin, den Addon zum aufsichtsrechtlichen Faktor von 3 im Value-at-Risk-Modell (VaR) zu minimieren. Gründe für einen Addon sind z.B. schlecht qualitätsgesicherte Marktdaten, die das Niveau des VaR verzerren. Schwächen des Risikomodells schlagen sich auch im Backtesting nieder, das einen Vergleich von aus der Risikoverteilung abgeleiteten Kennzahlen (Risikolauf) mit gegen den Markt gebenchmarkten Fair Values (MtM-Lauf) beinhaltet. Das Ampelkonzept des Backtestings bestimmt dann den Addon.
Vor diesem Hintergrund ist es offensichtlich, daß qualitätsgesicherte Marktdaten einen Mehrwert darstellen, weil sie Einschränkungen für das eigentliche Bankgeschäft, die aus einer unnötigen Kapitalbindung durch ein dysfunktionales VaR-Modell resultieren, auf ein Minimum reduzieren helfen.

Um dieses Thema systematisch abzuarbeiten, ist der Artikel folgendermaßen aufgebaut.

In Abschnitt 2 wird ein Marktdatenvalidierungsprozeß definiert, der zum Ziel hat, qualitätsgesicherte Marktdaten für die Risikomodelle zur Verfügung zu stellen. Dieser Prozeß ist in Abbildung 1 skizziert und besteht aus den folgenden Teilen, die im Folgenden näher erläutert werden:

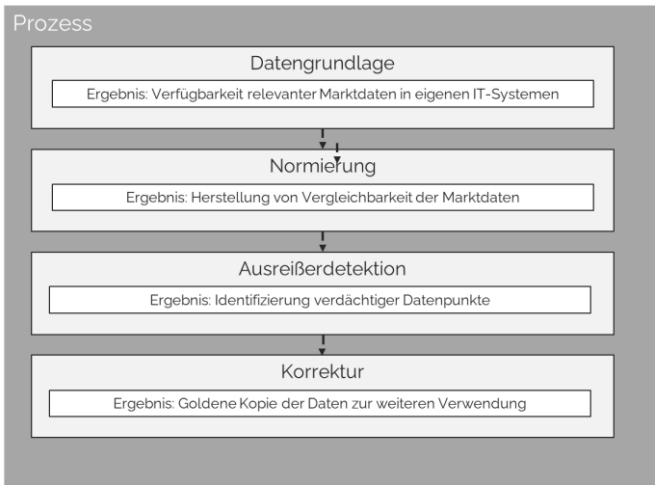

Abbildung 1: Struktur des Marktdatenvalidierungsprozesses

1. **Datengrundlage und Normierung**[2]: Dazu gehört zunächst die Festlegung des Umfangs des Datenabzugs. Gegebenenfalls wird ein und dieselbe fachliche Zeitreihe von unterschiedlichen Vendoren bezogen. Dann ist eine Normierung fachlich identischer Zeitreihen unterschiedlicher Vendoren zur Herstellung von Vergleichbarkeit erforderlich.

2. **Ausreißerdetektion:** Die Durchführung besteht aus zwei Teilschritten:

    a. Univariate Zeitreihenanalyse und
    b. Multivariate Zeitreihenanalyse.

    Diese Trennung erweist sich als nützlich, weil die Ausreißerdetektion und Behandlung auf ein Entscheidungsproblem führen, das mit Entscheidungsfehlern 1. bzw. 2. Art einhergeht (vgl. hierzu später Abschnitt 2.2.1).

3. **Korrektur:** Definition eines prototypischen Vorgehensmodells für die Bereinigung möglicher Ausreißer.

Der Artikel wird mit Abschnitt 3 beschlossen, der eine Zusammenfassung der wesentlichen Aussagen enthält.

[2] Auf den Datenabzug und die Normierung kann in diesem Artikel nur am Rande eingegangen werden, da die Darstellung einerseits von den konkreten Marktdaten abhängt und anderseits den Detaillierungsgrad einer technischen Feinspezifikation erfordern würde, die den Umfang dieses Artikels sprengen würde.

2 Marktdatenvalidierungsprozeß

2.1 Datengrundlage und Normierung

2.1.1 Marktdaten

Ein zentraler Aspekt für die Konsistenz der Daten aus einer oder verschiedenen Quellen ist ein einheitlicher Datenbezugs- oder Snapshotzeitpunkt[3]. Es sollte grundsätzlich ein konsistenter Datenabzug aus einer Quelle angestrebt werden (führender Anbieter).

In der Regel weisen die bereitgestellten Daten des führenden Anbieters Lücken oder Inkonsistenzen auf. Aus diesem Grund ist es sinnvoll, fachlich identische Zeitreihen von mehreren Anbietern (z.B. ICAP, Tullet Prebon, Tradition) zu sammeln, die größtenteils überschneidend angeboten werden, um diese bei fehlerhaften Daten des führenden Anbieters zu deren Vervollständigung verwenden zu können[4]. Dies gilt ebenso für den Fall komplementärer Verfügbarkeit von Marktdaten, die Ausnahmen von dem Paradigma des Datenabzug aus einer Quelle begründen.

Da die Marktdaten verschiedener Anbieter ggf. in unterschiedlichen Konventionen (Metadaten) bereitgestellt werden, ist für die spätere Vergleichbarkeit eine Normierung aller Marktdaten auf Soll-Metadaten erforderlich. Die Metadaten einzelner Zeitreihen müssen zuvor separat erhoben werden, da sie ggf. von Anbieter zu Anbieter differieren[5].

2.1.2 Ergänzende Information

Für die Einordnung der Qualität von Zeitreihen ist die Mitführung von zusätzlichen Informationen hilfreich. Ein Beispiel ist ein Liquiditätskennzeichen[6]. Dieses kann entweder von den Datenanbietern zur Verfügung gestellt werden oder es kann mittels eines Scoring-Verfahrens konstruiert werden. Weitere ergänzende Informationen sind vom konkreten Anwendungsfall abhängig.

2.2 Ausreißerdetektion

2.2.1 Grundsätzliche Überlegungen

Die Durchführung der Ausreißerdetektion wird im Rahmen des EOD-Prozesses durchgeführt und ist zeitscheibenbasiert. Unter Ausreißerdetektion wird zunächst

---

[3] Hier ist eine durchdachte technische Snapshot-Logik erforderlich, um Arbitragefreiheit - sogar innerhalb eines Datenlieferanten - zu gewährleisten.

[4] Dies ist ggf. mit zusätzlichen Lizenzkosten verbunden, da die Zeitreihe einmal vom führenden und einmal von einem alternativen Anbieter vorliegen muß.

[5] Die Normierung erfordert detaillierte fachliche Kenntnisse über die Marktkonventionen und Erfahrung in der Erstellung entsprechender technischer Mapping- und Umrechnungslogiken.

[6] Das Liquiditätskennzeichen kann z.B. auf Basis von Volumina tatsächlicher Transaktionen oder indikativer Preise und deren Bid-Ask-Spannen konstruiert werden.

nur eine Markierung eines neu angelieferten Datenpunktes als potentiell auffällig verstanden.

Zur Beurteilung eines neuen Datenpunktes ist eine Entscheidung erforderlich, die die zur Beurteilung des Datenpunktes verwendete historische Zeitreihe betrifft. Dies ist zunächst nur ein Thema für die univariaten Prüfungen, das dann später aber sinngemäß auch auf die multivariaten Prüfungen zu verallgemeinern ist.

Im univariaten Fall betrifft diese Entscheidung die Frage, wie bei der Detektion eines Ausreißers vorzugehen ist. Es existieren grundsätzlich zwei Alternativen:

i. Ausreißerdetektionen erfolgen auf Basis von Rohdaten, die keiner nachträglichen Manipulation unterworfen wurden[7].
ii. Ausreißerdetektionen erfolgen auf Basis von Daten, in denen vergangene Auffälligkeiten korrigiert wurden.

Wenn sich der Ausreißertest auf einen statistischen Hypothesentest (z.B. $4 \cdot \sigma$) stützt, dann ist der Unterschied der beiden Varianten durch die Natur des Fehlers bestimmt, den man bei der aktuellen Korrektur eines Datenpunktes tendenziell bereit ist einzugehen[8].

Die Nullhypothese geht immer von der Korrektheit der zu testenden Hypothese aus. Im Fall des $4 \cdot \sigma$-Tests ist die Nullhypothese, daß ein Wert in einem normalverteilten Sample mit einer Irrtumswahrscheinlichkeit von ca. 0.01% kleiner als $4 \cdot \sigma$ ist.

Es existieren zwei Varianten:

- Ein Fehler 1. Art liegt vor, wenn bei einem Hypothesentest die Nullhypothese zu Unrecht verworfen wird. Das entspricht der Situation, wo ein potentieller Ausreißer korrigiert wird, obwohl er ein echtes Event darstellt. Dieses Ergebnis wird begünstigt, wenn man mit Zeitreihen arbeitet, in denen Auffälligkeiten aus der Vergangenheit bereits wegdefiniert wurden, so daß die Volatilität der Referenzzeitreihe sinkt (Variante ii).
- Ein Fehler 2. Art liegt vor, wenn bei einem Hypothesentest die Nullhypothese zu Unrecht beibehalten wird. Dies entspricht der Situation, wo ein potentieller Ausreißer deshalb nicht erkannt wird, weil die Volatilität der Referenzzeitreihe aufgrund der Beibehaltung historischer Ausreißer tendenziell steigt (Variante i).

Ein Fehler 1. Art geht tendenziell mit der Verwendung der Alternative ii einher während ein Fehler 2. Art tendenziell mit der Verwendung der Alternative i einhergeht. Priorität wird der Vermeidung von Fehlern 1. Art eingeräumt (d.h. keine

[7] Diese Forderung entspricht einer idealisierten Situation. Sie unterstellt, daß die historischen Daten keine offensichtlichen Datenfehler (z.B. Kommaversatz) enthalten. Wir machen hier zunächst aus didaktischen Gründen diese Vereinfachung und gehen in Abschnitt 2.4 nochmals auf die allgemeinere Situation ein.

[8] Nicht alle Tests sind notwendig statistischer Natur. Dennoch kann das hier vorgestellte Prinzip als ein leitendes Prinzip verwendet werden, welches in den anderen Fällen immer noch im Sinne einer Analogie angewendet werden kann.

Volatilitätsunterschätzung!), weil diese im Zweifel mit einer Risikounterschätzung verbunden sind. Aus diesem Grund wird die Alternative i bevorzugt.

Bisher haben wir ausschließlich die univariate Ausreißerdetektion betrachtet. Wenn man auch die multivariate Ausreißerdetektion mit in die Betrachtung einschließt, ergibt sich eine zusätzliche Möglichkeit zur Hebung der Datenqualität. Der typische Use Case ist eine im Rahmen der univariaten Tests unentdeckt gebliebene Auffälligkeit in einer Datenzeitreihe, die im Zuge der multivariaten Prüfung dann doch noch erkannt und korrigiert werden kann. Da die univariaten Test so konstruiert sind, daß sie tendenziell keine Fehler 1. Art begehen (korrekte Werte werden als auffällig markiert), handelt es sich bei den verbliebenen Fehlern in der Mehrzahl um Fehler 2. Art, d.h. der univariate Test hat einen falschen Wert nicht als auffällig markiert. Es wird aber auch noch ein wichtiger Fall eines Fehlers 1. Art durch die multivariaten Tests mitabgedeckt. In genau diese Lücke greifen die multivariaten Tests ein.

Diese Rangfolge der Behandlung von Fehlerarten begründet auch noch einmal nachträglich die logische Zerlegung in univariate und multivariate Tests und die Reihenfolge ihrer Ausführung.

Im Vorgriff auf die in Abschnitt 2.3 zu behandelnde Datenkorrektur ergibt sich die Notwendigkeit die Zeitreihen in mindestens drei Fassungen vorzuhalten:

a) als Rohdaten (Bezug),
b) als (ggf. mehrfach) markierte Zeitreihen (Ausreißerdetektion) und
c) als korrigierte Zeitreihen (goldene Kopie).

2.2.2 Univariate Prüfungen

Der Fall der univariaten Prüfungen ist in der Bankpraxis weitestgehend standardisiert, so daß wir uns hier lediglich auf eine Aufzählung der typischsten Tests beschränken. Diese umfassen i.d.R. die folgenden Tests:

- Zeitreihe besitzt Datenlücken[9],
- Zeitreihe hat Vorzeichenwechsel[10],
- Zeitreihe hat Werte gleich Null,
- Zeitreihe seit N Tagen konstant,
- Zeitreihe besitzt Werte größer als $n \cdot \sigma$,
- Zeitreihe bei alternativem Anbieter.

Grob kann man die Tests in technische Vollständigkeitstests, fachliche Plausibilitätstests und statistische Tests (Hypothesentests) unterteilen, wobei der Übergang z.T. fließend ist. So kann man den Test auf konstante Werte durchaus

[9] Die Vervollständigung von Datenlücken wird erst im Abschnitt 2.3 besprochen. Allerdings ist für die Berechnung von statistischen Kennziffern eine vollständige Datenreihe erforderlich. Für unsere Zwecke wird die Berechnung der statistischen Kennziffern auf verjüngten Zeitreihen, d.h. um die fehlenden Daten verkürzten Zeitreihen, durchgeführt. Die Returns können ggf. noch um die Länge der Bezugsperiode korrigiert werden. Allgemeinere Methoden finden sich in [2].
[10] Verallgemeinerung des klassischen Tests „Zeitreihe hat negative Werte“.

im Sinne eines Run-Tests interpretieren und damit einem Hypothesentest zugänglich machen.

Für uns ist jedoch die Verbindung der Ergebnisse der statistischen univariaten Ausreißertests (z.B. n•σ) mit den Ergebnissen von multivariaten Tests interessanter, der wir unsere Aufmerksamkeit im folgenden Abschnitt zuwenden werden[11].

2.2.3 Multivariate Analysen

2.2.3.1 Auswahl Vergleichszeitreihen

Um multivariate Analysen zur Qualitätssicherung von Marktdatenzeitreihen (Referenzzeitreihe) durchzuführen, müssen zunächst Vergleichszeitreihen definiert werden.

Die natürlichen Vergleichszeitreihen sind Marktdatenzeitreihen, die eine fachliche Abhängigkeit zu der zu untersuchenden Referenzzeitreihe aufweisen. Da es sich dabei ggf. um eine Vielzahl von infrage kommenden Vergleichszeitreihen handelt, werden diese in Gebinden zusammengefaßt, die mit einer empirischen Korrelationsmatrix (lineares Abhängigkeitsmodell) einhergehen.

Eine sinnvolle Vergleichszeitreihe ist eine Zeitreihe, die eine möglichst hohe positive[12] Korrelation mit der Referenzzeitreihe aufweist.

Im Fall, daß keine sinnvollen Vergleichszeitreihen festgelegt werden können, reduziert sich der multivariate Test genau auf den entsprechenden univariaten ‚n•σ'-Test.

2.2.3.2 Mahalanobis und bivariate Subtests

Die Identifikation von multivariaten Ausreißern kann dann sowohl durch eine automatisierte Untersuchung der Zeitreihen mit statistischen Methoden als auch über graphische Analysen erfolgen. Die statistischen Methoden verwenden z.B. den Mahalanobis-Abstand[13] [1]:

$$d^2(\vec{x}, \vec{a}) = (\vec{x} - \vec{a})^T \Sigma^{-1} (\vec{x} - \vec{a}).$$

Hierbei bezeichnet Σ die Kovarianzmatrix und $\vec{x}$ den Abstand vom Schwerpunkt $\vec{a}$ der Punktwolke. Die Streuellipsen der Punktwolke sind die Menge der Punkte mit konstantem Mahalanobis-Abstand $d^2(\vec{x}, \vec{a}) = k^2$. Sie entsprechen bei elliptischen Verteilungen den Höhenlinien gleicher Wahrscheinlichkeitsdichte (Isodensiten) und sind damit eine direkte Verallgemeinerung der Varianz aus dem eindimensionalen Fall.

Man kann die Systematik, ab wann ein Datenpunkt als Ausreißer zu klassifizieren ist, dann z.B. an den Mahalanobis-Abständen $k_1, \dots, k_n$ festmachen, die zu den gleichen Wahrscheinlichkeiten p_i mit $i = 1, \dots, n$ wie im 1-dimensionalen Fall, d.h.

---

[11] Das Thema Data-Imputation wird hier bewußt abgegrenzt. Wir werden im Abschnitt 2.4 aber noch einmal kurz darauf zurückkommen.

[12] Grundsätzlich ist aus theoretischer Sicht auch eine hohe negative Korrelation geeignet. In der Praxis wird aber der andere Fall bevorzugt.

[13] Es existieren feinere statistischen Methoden, die aber zu technisch sind, um im Rahmen dieser Präsentation vorgestellt zu werden.

95%, 99% etc. p.p. gehören. Der elementare Anwendungsfall besteht nun darin, nach der verallgemeinerten $'n \cdot \sigma'$-Logik potentielle Ausreißer zu identifizieren.

Grundsätzlich können auch korrekte Datenpunkte zu einer Verletzung des $'n \cdot \sigma'$-Kriteriums führen. Um diesen Fall abzutrennen, ist es sinnvoll, die zu prüfende Zeitreihe nicht nur in einem multivariaten Test über das gesamte Gebinde, sondern auch in bivariaten Tests gegen jede einzelne Vergleichszeitreihe aus dem Gebinde zu prüfen. Dadurch kann das Ausreißersignal aus dem multivariaten Test weiter analysiert werden[14].

Die Grundidee dieser Korrelationstests beruht darauf, daß ein potentieller Ausreißer in Bezug auf die Marginalverteilung einer Zeitreihe genau dann plausibilisiert ist, wenn sich das gleiche Ausreißersignal in der entsprechenden Marginalverteilung der Vergleichszeitreihe ebenso nachweisen läßt. Dies führt zu einer Verringerung der Entscheidungsfehler 1. Art aus der Perspektive der univariaten statistischen Tests[15].

Umgekehrt kann man durch eine inverse Anwendung der Methode auch Entscheidungsfehler 2. Art identifizieren[16]. Das liegt daran, daß ein Ausreißer in einem Zeitreihenpaar einen sensitiven Einfluß auf die empirische Korrelation des Zeitreihenpaars besitzt. Selbst wenn der Ausreißer nicht groß genug ist, um in dem Volatilitätstest auffällig zu werden, kann er noch durch seinen negativen Einfluß auf die empirische Korrelation detektiert werden.

Zu Illustrationszwecken ist dieser Effekt in Abbildung 2 am Beispiel einer bivariaten Verteilung zweier korrelierter Zeitreihen dargestellt: Die Korrelation kann – grob gesagt – als die Verdrehung der Hauptachsen der Ellipse gegenüber dem Koordinatensystem verstanden werden. Es ist unmittelbar einsichtig, daß der Datenpunkt in der linken oberen Ecke

a) einen potentiellen Ausreißer darstellt und
b) einen wesentlichen Einfluß auf die Hauptachsen der Punktwolke hat.

Da die Bestimmung der Hauptachsen etwas Algebra erfordert, sind in Abbildung 2 nur die mit Excel-Bordmitteln bestimmbaren Regressionsgeraden für die beiden Punktwolken mit/ohne Ausreißer eingezeichnet. Aber auch hier zeigt sich der Effekt des Ausreißers in aller Deutlichkeit in einer kompletten Umdrehung des Vorzeichens des Anstiegs der Regressionsgeraden, einem leicht zu identifizierenden Event[17].

---

[14] Stammt der multivariate Ausreißer aus nur einer univariaten Zeitreihe, so läßt sich diese Zeitreihe identifizieren. Dies erfolgt durch Sortierung der Zeitreihen nach der Anzahl von Korrelationsausreißern gegen die verschiedenen bivariaten Tests.

[15] Als Spezialfall fachlich abhängiger Zeitreihen wird auch der Fall der Verfügbarkeit der Zeitreihe bei einem alternativen Anbieter betrachtet. Das führt dazu, daß ein vermeintlicher Volaausreißer durch einen Alternativanbieter bestätigt werden kann.

[16] Die direkte Anwendung geht von einer korrekt geschätzten Korrelationsmatrix aus. Die inverse Anwendung prüft den Impact eines Datenpunktes auf den Schätzer der Korrelation.

[17] Die Methode der bivariaten Korrelationsausreißer ist bereits für sich genommen eine hochsensitive Methode zur Identifikation von potentiellen Ausreißern, die mit den Mitteln der

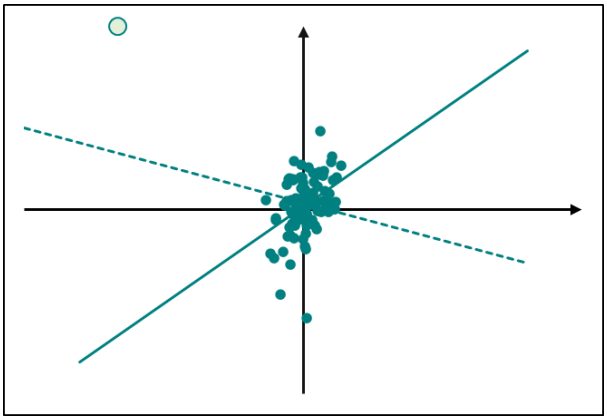

Abbildung 2: Illustration des Effekts eines Ausreißers auf die Regressionsgeraden einer Datenwolke

Die bivariaten Verteilungen ergeben sich dabei aus der multivariaten Verteilung durch Projektion auf Paare korrelierter Variablen. Die Zeitreihe, die den Ausreißer produziert, ergibt sich wieder durch Analyse der Anzahl der bivariaten Signale (vgl. Fußnote 14). War der Datenpunkt der so identifizierten Zeitreihe in der univariaten Analyse noch unauffällig, so steigt die Wahrscheinlichkeit, daß es sich hierbei um einen Korrelationsausreißer handelt, der einem Fehler 2. Art aus univariater Perspektive entspricht, mit der Anzahl der bivariaten Signale sehr schnell an.

Eine sinnvolle Dimension für multivariaten Tests ist demzufolge 3. Bei Gebinden aus Zeitreihen, die z.B. Zinskurven darstellen, entspricht das z.B. den beiden nächsten Nachbarn.

In diesem Sinne kann der multivariate Test und die Menge seiner ererbten bivariaten Subtests verwendet werden, um ein Kreuzverhör bzgl. der Aussagen der univariaten Tests durchzuführen[18].

2.3 Datenkorrektur

2.3.1 Verwendung der Ausreißersignale

Dieser Abschnitt stellt einfache Prinzipien vor, denen der manuelle Korrekturprozeß zur finalen Erstellung einer Goldenen Kopie von Marktdatenzeitreihen folgt.

Das Resultat der Ausreißeranalysen aus dem vorigen Abschnitt 2.2 ist eine Liste von potentiellen Ausreißern. Diese stellen die Ausgangslage dar, von der die manuelle Pflege der neu in die Zeitreihen aufzunehmenden Marktdatenpunkte startet.

Der manuelle Pflegeprozeß läßt sich dabei weitestgehend vorautomatisieren. Dies geht sogar so weit, daß die Entscheidungen zu einzelnen potentiellen Ausreißern

---

univariaten Analyse überhaupt nicht zu erkennen sind. Aus dieser Perspektive rangiert sie gleichrangig zu den univariaten Tests.

[18] Auf der Ebene der multivariaten Tests wiederholt sich das Problem der Entscheidungsfehler 1. und 2. Art analog. Durch den Zuwachs an korrelierten Informationen wächst die Entscheidungssicherheit jedoch stetig an. Wir führen deshalb keine Iteration über diese nachgelagerten Entscheidungen mehr durch, sondern unterstellen auf diesem Niveau bereits Sicherheit.

mit allen fachlichen Argumenten und inklusive der Datenvervollständigungs-vorschläge schablonenartig vorbereitet werden können, so daß der manuelle Prozeß praktisch nur noch in einem Review von mundgerecht vorbereiteten Entscheidungen besteht. Dies reduziert den manuellen Aufwand auf ein Minimum bei Beibehaltung einer hohen fachlichen Qualität.

Dazu wird im Kern der Entscheidungsbaum aus Abbildung 3 verwendet, der noch einmal die verschiedenen Filtertypen mit den möglichen Ergebnissen des Korrekturprozesses in Zusammenhang bringt.

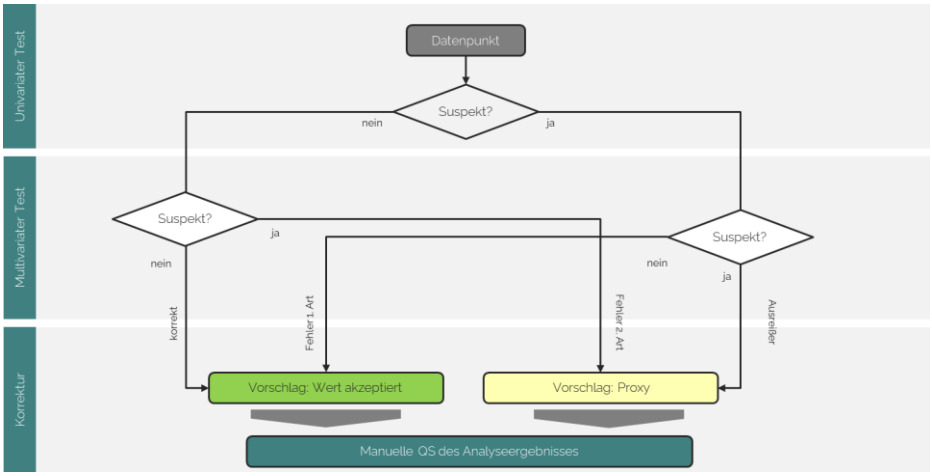

Abbildung 3: Entscheidungsbaum zur Erzeugung eines qualitätsgesicherten Datenpunktes

Beim Durchlaufen dieses Baumes sind grundsätzlich nur folgende Fälle möglich:

1. Datenpunkt unauffällig bzw.
2. Datenpunkt in univariaten oder multivariater Prüfungen auffällig, wobei sich drei Unterausprägungen ergeben:
    a. Der Datenpunkt ist nur in der univariaten aber nicht in der multivariaten Analyse auffällig (potentieller Fehler 1. Art),
    b. Der Datenpunkt ist in der multivariaten (insbesondere den bivariaten Subtests) aber nicht in der univariaten Analyse auffällig (potentieller Fehler 2. Art).
    c. Der Datenpunkt ist sowohl in der univariaten als auch in der multivariaten Analyse auffällig (potentieller Datenfehler).

Der 1. Fall und der Fall 2.a, der der Identifikation eines Entscheidungsfehlers 1. Art aus der univariaten Analyse enthält, führen i.d.R. zur Akzeptanz des Datenpunktes. Der Fall 2.b, enthält die Identifikation eines Entscheidungsfehlers 2. Art aus der univariaten Analyse. Er erfordert ggf. die Erzeugung eines Proxies, der so zu wählen ist, daß durch den Proxy die ‚korrekte' empirische Korrelation der Zeitreihen erhalten wird (vgl. Abschnitt 2.3.2 weiter unten). Der Fall 2.c führt zur Ermittlung eines Proxy-Wertes mit Standardmethoden.

2.3.2 Datenvervollständigung und Proxyerzeugung

Die entsprechenden Proxymechanismen sind schematisch in Abbildung 4 zusammengefaßt.

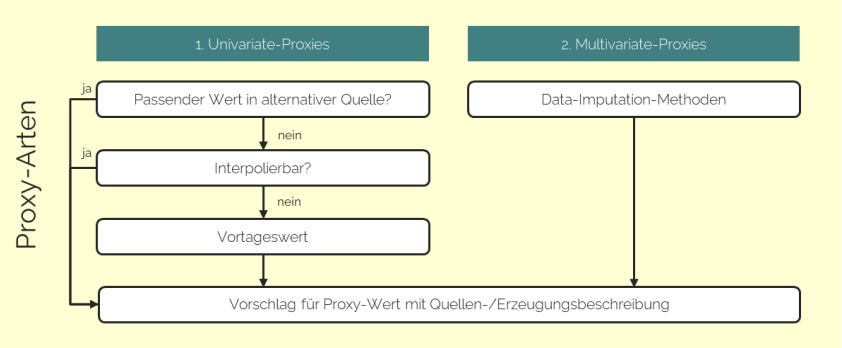

Abbildung 4: Übersicht über Proxymechanismen

In der Regel werden die univariaten Proxies verwendet, um einen geeigneten Schätzer für einen auffällig gewordenen Datenpunkt zu erzeugen. Dabei ist immer zu prüfen, ob der verwendete Proxy die identifizierten Auffälligkeiten in den beiden Dimensionen Volatilität und Korrelation angemessen korrigiert.

Eine besondere Rolle kommt dem multivariaten Proxy zu, die immer dann angewendet werden muß, wenn kein geeigneter univariater Proxy zur Verfügung steht, oder der univariate Proxy nicht in der Lage ist, das Auffälligkeitsmuster überzeugend aufzulösen.

Ein wichtiger Anwendungsfall sind Bondportfolien, die aufgrund unterschiedlicher Liquidität mal mehr oder weniger vollständige Historien aufweisen (vgl. z.B. [3]).

Das typische Beispiel für einen multivariaten Proxy ist die Methode der C(onditonal)E(xpectations).

Die CE-Methode produziert für auffällige oder fehlende Daten, die nicht durch univariate Proxies korrigiert werden konnten, einen Proxy, der auf einer Regression gegen beobachtbare Daten basiert. Die Regression ist dabei durch die historisch beobachtete Korrelation[19] der abhängigen Größen bestimmt. Sie berechnet sich als konditionaler Erwartungswert und hat bei der Verwendung einer um Null zentrierten 2-dimensionalen Normalverteilung z.B. die Gestalt:

$$\mathbb{E}_X[Y|X = x] = \rho \cdot \frac{\sigma_Y}{\sigma_X} \cdot x.$$

Hier bezeichnen X, Y zwei korrelierte Zufallsvariablen und σ_X und σ_Y ihre Volatilitäten und ρ die historische Korrelation. Durch die Verwendung höherdimensionaler Verteilungen läßt sich die Qualität dieses Proxies ggf. steigern.

Die volle Wirkung entfaltet die CE-Methode im Zusammenspiel mit dem EM-Algorithmus [2].

[19] Zur Berechnung der historischen Korrelation sind geeignete statistische Methoden anzuwenden, da die historische Korrelation nach dem bisher Gesagten auch durch ein echtes Event verzerrt werden kann (vgl. z.B. [2]).

2.4 Epilog

In diesem Abschnitt soll noch auf ein paar wichtige Spezialthemen eingegangen werden, die im Zuge der gewählten Darstellung zu kurz gekommen sind. Diese betreffen insbesondere die Rohdatenzeitreihe, die zur Ermittlung von potentiellen Ausreißern verwendet wird.

In den grundlegenden Überlegungen von Abschnitt 2.2.1 haben wir den Standpunkt vertreten, daß die Ausreißerdetektion auf den Rohdaten aufzusetzen hat. Hier wollen wir diesen Standpunkt in zwei Punkten nachschärfen:

1. Initiale Validierung der Rohdaten: Wird eine Zeitreihe neu in den Marktdatenbezug aufgenommen, so ist eine initiale Validierung der relevanten Historie der Rohdatenzeitreihen durchzuführen. Diese bedient sich grundsätzlich der zuvor beschriebenen Methoden, hat aber das Ziel, offensichtliche Datenfehler zu bereinigen. Das bedeutet, daß die Validierung jedes einzelnen auffälligen Datenpunktes durch den Abgleich gegen eine Referenzquelle bzw. die Begründung, welches Event zu der potentiellen Auffälligkeit geführt hat, durchzuführen ist.
2. Dasselbe gilt für im Zeitverlauf nachträglich identifizierte Datenfehler.

In diesem Sinne ist die sehr strikte Bezugnahme auf die Rohdaten als Basis der statistischen Validierungen zugunsten einer Unterteilung in eine originale Rohdatenzeitreihe und eine um eindeutig identifizierbare Datenfehler bereinigte abgeleitete Rohdatenzeitreihe zu modifizieren. Alle Modifikationsschritte sind einem Audit Trail zu unterwerfen, um diese berechtigten Manipulationen später nachverfolgen zu können. Die Statistik setzt dann tatsächlich auf dieser zweiten Variante der Zeitreihe auf.

In der in diesem Artikel gewählten Darstellung haben wir den Prozeß der Ausreißerermittlung und der ggf. nötigen Datenvervollständigung als zwei separate Prozeßschritte dargestellt. Es liegt in der Natur der Sache, daß die beiden Prozeßschritte miteinander verknüpft sind. Es existieren etablierte Methoden, die die beiden Schritte miteinander verbinden. Prominent zu nennen ist hier der EM-Algorithmus [2]. Eine Anwendung auf praktische Probleme des Bankbetriebs findet sich z.B. in [3].

3 Zusammenfassung

Eine zentrale Voraussetzung für ein möglichst aussagekräftiges Risikomodell ist neben der verwendeten Methodik insbesondere eine qualitativ hochwertige Zeitreihenbereinigung. Dies betrifft neben dem Niveau der Volatilität insbesondere auch das Thema der implizit in den Zeitreihen kodierten Korrelationen von Risikofaktoren. So würde sich eine Risikodiversifikation durch eine negative und eine Risikokonzentration durch eine positive Korrelation der relevanten Zeitreihen von Risikofaktoren manifestieren.

Wie in diesem Artikel ausgeführt, kann ein Ausreißer in einer Zeitreihe einen signifikanten Einfluß auf die Korrelation zu anderen Zeitreihen haben. Eine Nichtbeachtung der Auswirkung eines Ausreißers auf die Korrelationen im Rahmen der Datenkorrektur führt dann zu einer subtilen Fehleinschätzung des Risikos. Im Extremfall können sich die Verhältnisse von Diversifikation und Konzentration sogar vollständig verkehren.

Um diesem Phänomen geeignet Rechnung zu tragen, sind in einem Marktdaten-validierungsprozeß neben den univariaten Qualitätsprüfungen, welche die Auswirkungen auf die Volatilität identifizieren aber blind gegenüber der Korrelation zu anderen Zeitreihen sind, auch multivariate Qualitätsprüfungen erforderlich, die so auszulegen sind, daß eine unplausible Veränderung des fachlichen Gehalts der Korrelation von Paaren von Zeitreihen sicher erkannt wird. Als Folge können Ausreißer, die den oben beschriebenen Effekt der Korrelationsverzerrung hervorrufen, identifiziert und im Korrekturprozeß behandelt werden. Die Korrelation kann aber auch umgekehrt zur Bestätigung von bzgl. der Volatilität auffällig gewordenen echten Events verwendet werden.
Aus einer klassifizierenden Perspektive, erlauben die multivariaten Tests die Identifizierung von Fehlern 1. und 2. Art aus den univariaten Tests.

Es sollte deutlich geworden sein, daß gerade das Ziel der Adäquanz des Risikomodells die Berücksichtigung von multivariaten Ausreißertests somit zwingend erfordert. Die systematische Verwendung der multivariaten Tests wirkt sich positiv auf die Qualität der Marktdaten bei gleichzeitiger Reduktion der Kosten aus. Viele der nötigen Vergleiche gegen einen alternativen Anbieter können vermieden werden, da die gleiche Information bereits aus den Daten des führenden Anbieters abgeleitet werden kann. Gerade für kostenbewußte Anwender ist die Verwendung algorithmischer Lösungen auf Basis multivariater Analysen somit preisgünstiger als die mit Lizenzkosten verbundenen Buy-Lösungen.

Da fachlich adäquater, reduzieren qualitätsgesicherten Marktdaten im Ergebnis den operativen Aufwand zur Kommentierung der Ergebnisse des Risikomodells im täglichen Linienprozeß. Ebenso tragen sie zur Erreichung des wirtschaftlichen Ziels bei, Einschränkungen für das Kerngeschäft einer Bank, die aus einer unnötig hohen Eigenmittelbindung durch das Risikomodell resultieren, auf ein Minimum zu reduzieren.

4 Literatur

[1] P. C. Mahalanobis, "On the generalised distance in statistics". In: *Proceedings of the National Institute of Science of India*. Band 2, Nr. 1, 1936, S. 49–55

[2] A. P. Dempster, N. M. Laird, D. B. Rubin, "Maximum-Likelihood from incomplete data via the EM algorithm*", Journal of the Royal Statistical Society.* 1977, S. 1–38

[3] T. Siegl, P. Quell „Modelling Specific Interest Rate Risk with Estimation of Missing Data", *Appl. Math. Fin.* 12(3):283-309

Danksagung: Die Autoren bedanken sich bei Tilman Wolff-Siemssen für eine kritische Durchsicht des Textes und Anregungen zur Verbesserung der Darstellung.