DWDS Wortverläufe

Bryan Jurish
jurish@bbaw.de

Berlin-Brandenburgische Akademie der Wissenschaften
31. Mai 2016

Überblick

Ausgangslage

- Korpora & Frequenzen
- Histogramme & Zeitreihen
- Desiderata

Implementierung

- Zählen & Aggregieren: `count(...) #by[...]`
- Graphematische Varianz & lexikalische Identität
- Glättung & Ausreißererkennung

Zusammenfassung

Korpora & Frequenzen

DWDS Korpora

- 43 verschiedene Korpora
- 9.4 Milliarden Tokens in 553 Millionen Sätzen
- Automatisch linguistisch annotiert & mit DDC/D* indiziert
- Zeitraum: 1600–2016 *. . . jedoch nicht gleichverteilt!*

(Wort)frequenzen

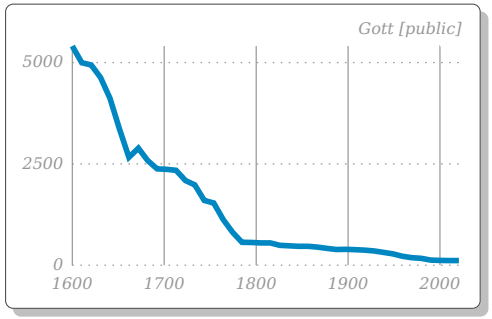
- Indikator für (lexikalische) Eigenschaften, e.g. Zugriffsgeschwindigkeit
- *Nicht konstant* über die Zeit
- Diachrone Veränderungen können auf lexikalische Prozesse hindeuten

Histogramme & Zeitreihen

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Histogramm

- *Häufigkeitsverteilung* nach unabhängiger Kategorie (x)

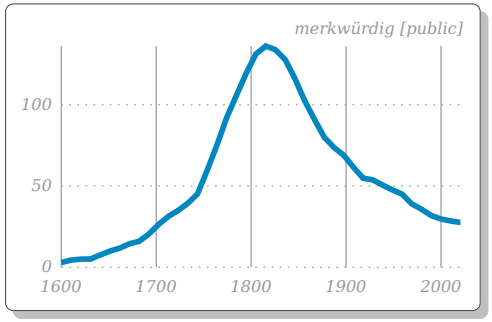


Zeitreihe

- *Zeitabhängige* Folge von Datenpunkten (y)

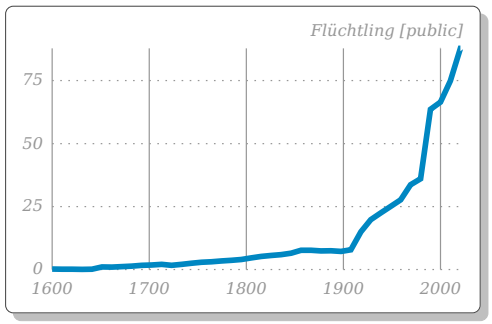
Wortverläufe $\sim$ Zeitreihenhistogramme

- Histogramm-Kategorien (x-Achse): Zeitintervalle
- Zeitreihen-Datenpunkte (y-Achse): Häufigkeiten



Graphische Darstellung

- Zweidimensionaler Plot
- Datenpunkte werden per Strecken verbunden
  - (lineare) Interpolierung $\sim$ zeitliche Kontinuität



Desiderata

Datengewinnung

- Korpus-abhängige Frequenzen — Systemintegrierung
- Lemma-basierte Frequenzen — {*Arzt, Arztes, Ärzte, Ärzten*} ∼ *Arzt*
- Schreibweisentoleranz — {*Theil, Theyl, Tayl, . . .*} ∼ *Teil*
- Flexibilität (für erfahrene Anwender)
  - Reguläre Ausdrücke — `/panzer$/, /^Hilfs/, . . .`
  - Phrasensuche — *mit nichten, in der Tat, . . .*
  - PoS-Filter — *einen*$_{\mathrm{DET|V}}$, *Leben*$_{\mathrm{N|V}}$, *Würde*$_{\mathrm{N|V}}$, . . .

Darstellung

- Direkte Verknüpfung mit Korpustreffern — Interoperabilität
- Sortiert nach Textklasse — *Wissenschaft*, *Belletristik*, *Zeitung*, . . .
- Flexible Granularität der Zeitachse — Jahr, Dekade, Jahrhundert, . . .
- Intuitive Glättung der graphischen Darstellung — (spärliche, heterogene Daten)
- Erkennung & Herausfiltern von Ausreißern — (Korpusartefakte)
- Exportfunktion / externes API — Weiternutzung

Implementierung

Zählen und Aggregieren

Idee: Datengewinnung per Suchmaschine

- Alle DWDS Korpora sind bereits mit der DDC Suchmaschine indiziert
- DDC unterstützt Lemma-Suche, Phrasensuche, PoS-Filter, usw.
- Interoperabilität $\rightsquigarrow$ ohne weiteren Aufwand

Herausforderung: Effizienz

- Einzelabfragen für jeden Datenpunkt $\rightsquigarrow$ *zu langsam*

Lösung: neuer DDC `count()`-Operator

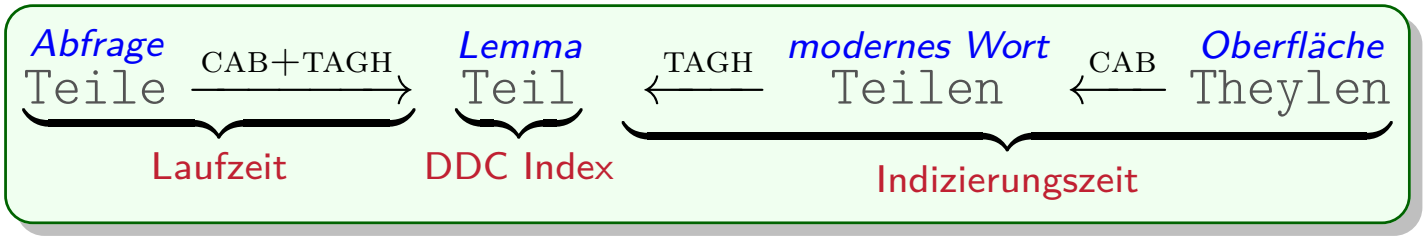
- Bis *1600x schneller* als Einzelabfragen *(DTA+DWDS, Einzeljahre)*
- Abstrakte Implementierung vielseitig einsetzbar *(e.g. zur Kollokatensuche)*
- Zeitreihe für Abfrage Q mit Intervall S nach primärer Textklasse:

```
count(Q) #by[date/S, textClass~s/:.*//]
```

Graphematische Varianz

Ziel: Schreibweisentolerante Suche für historische Korpora

- Suche nach *moderner Wortform* sollte auch alle *historischen Varianten* finden



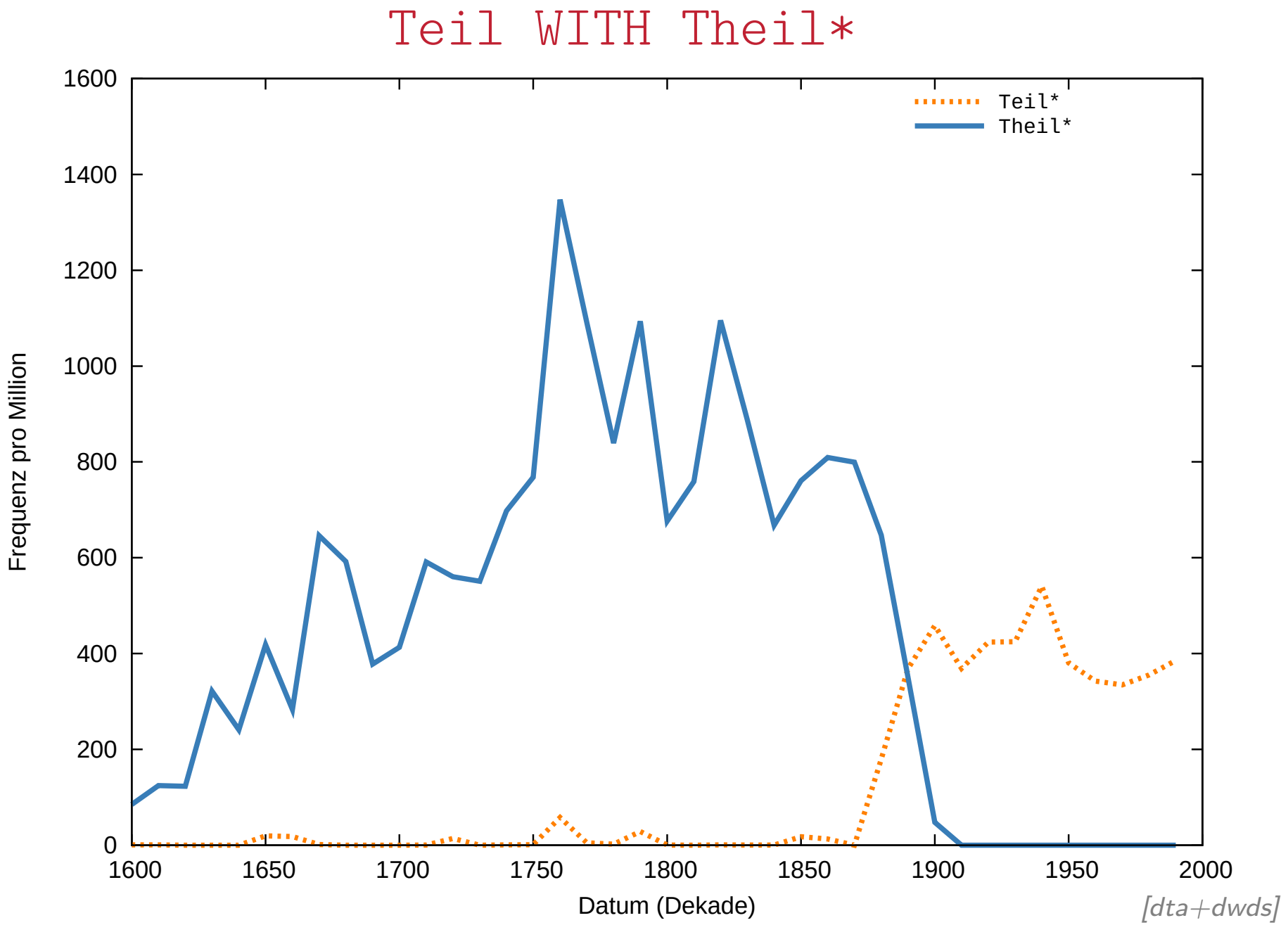
Methode: CAB

(Jurish, 2011)

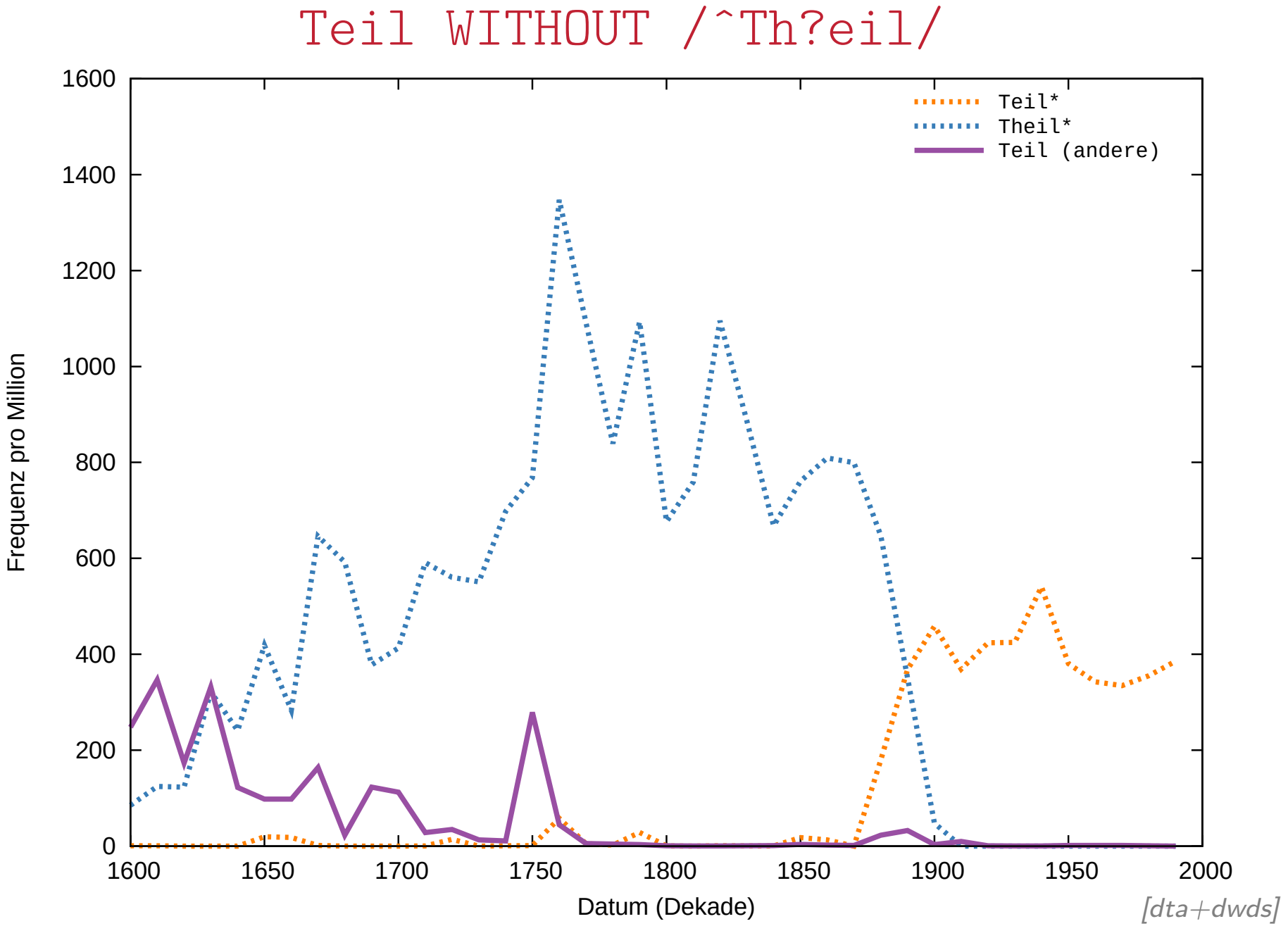
- "Cascaded Analysis Broker" zur robusten Analyse historischer Texte
- Liefert (modernen) *"kanonischen Kognaten"* zur (historischen) Wortform
  - Transliterierung, Phonetisierung, & gewichteter "rewrite"-Regelsatz
  - Satzweise Optimierung mittels dynamischen Hidden Markov Modell
  - weitere automatische Analyse kann sich auf *Normalformen* beziehen
- **Korpusabhängig:** nur aktiv für historische Korpora (DTA, Dingler)
- **Optional:** kann mit direkter Suche (Regex, @-Operator) umgangen werden

Graphematische Varianz: der Fall *"Teil"*

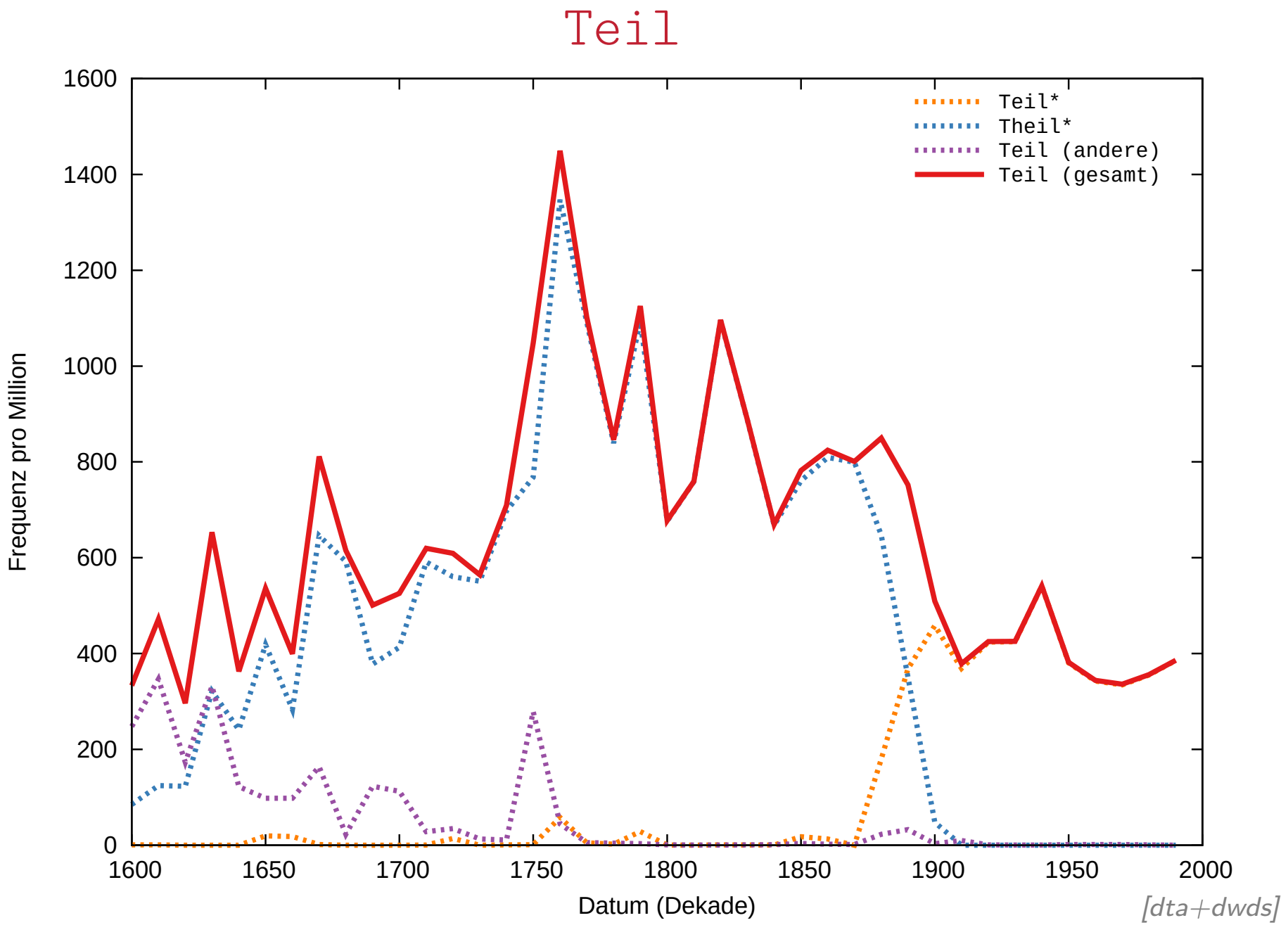
Graphematische Varianz: der Fall *"Teil"*



Graphematische Varianz: der Fall *"Teil"*



Graphematische Varianz: der Fall *"Teil"*

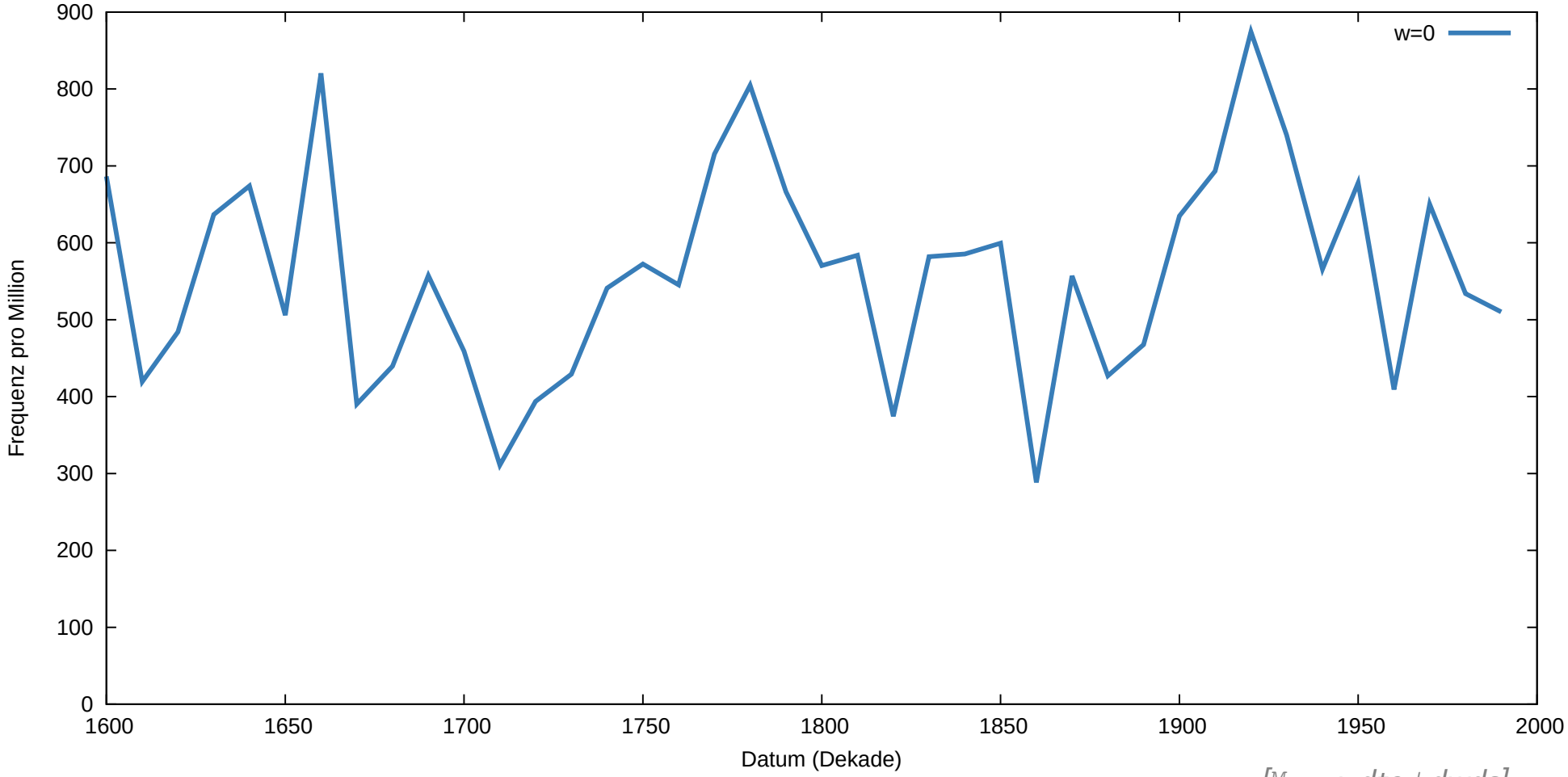


Glättung: lokale Schwankungen

- **Problem:** gemessene Frequenzen schwanken lokal ("zick-zack")

Glättung: lokale Schwankungen

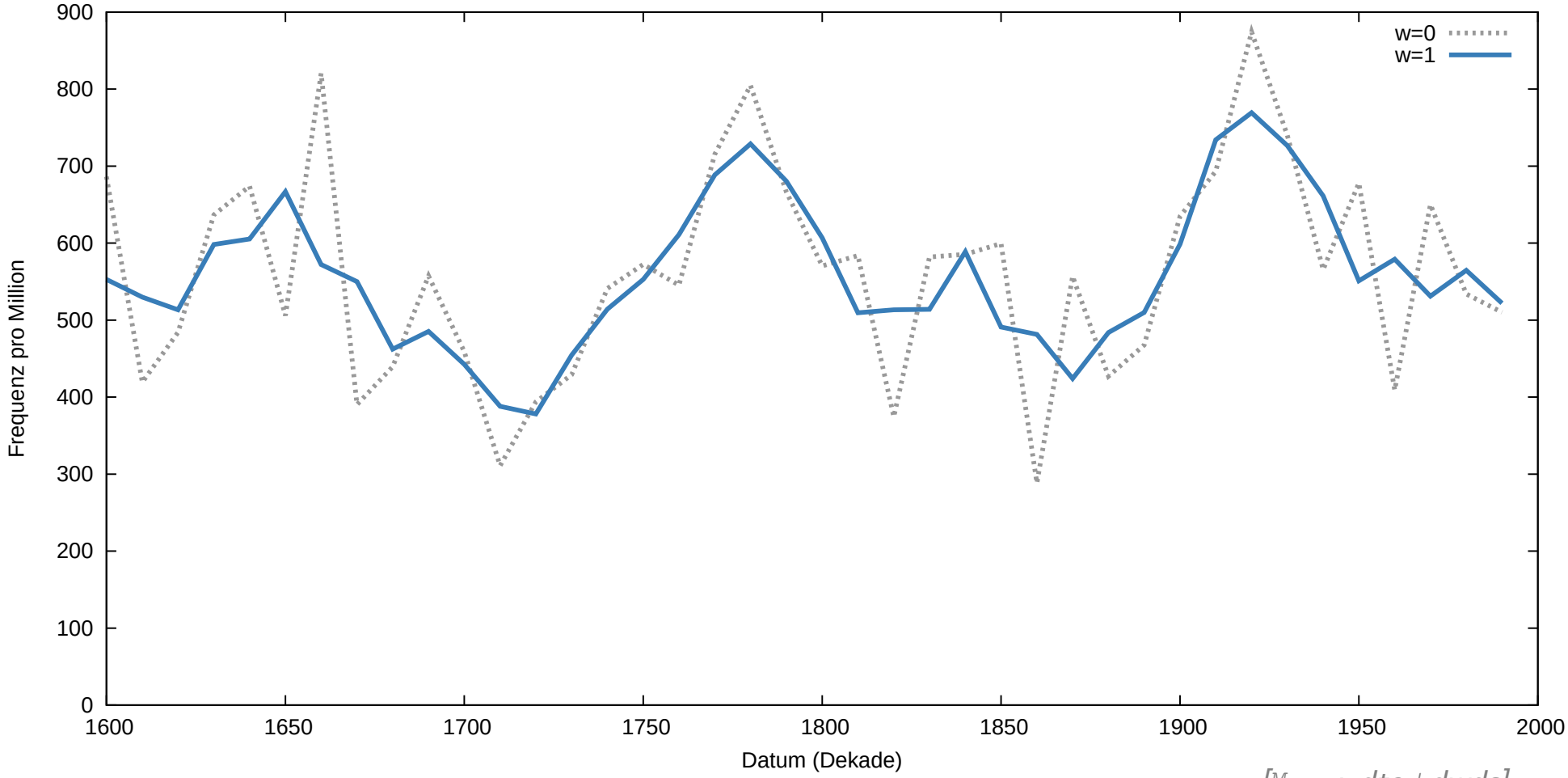
- **Problem:** gemessene Frequenzen schwanken lokal ("zick-zack")
- **Ursache:** (zu) kleine Intervalle, spärliche Daten, "Rauschen"



[Mann: dta+dwds]

Glättung: lokale Schwankungen

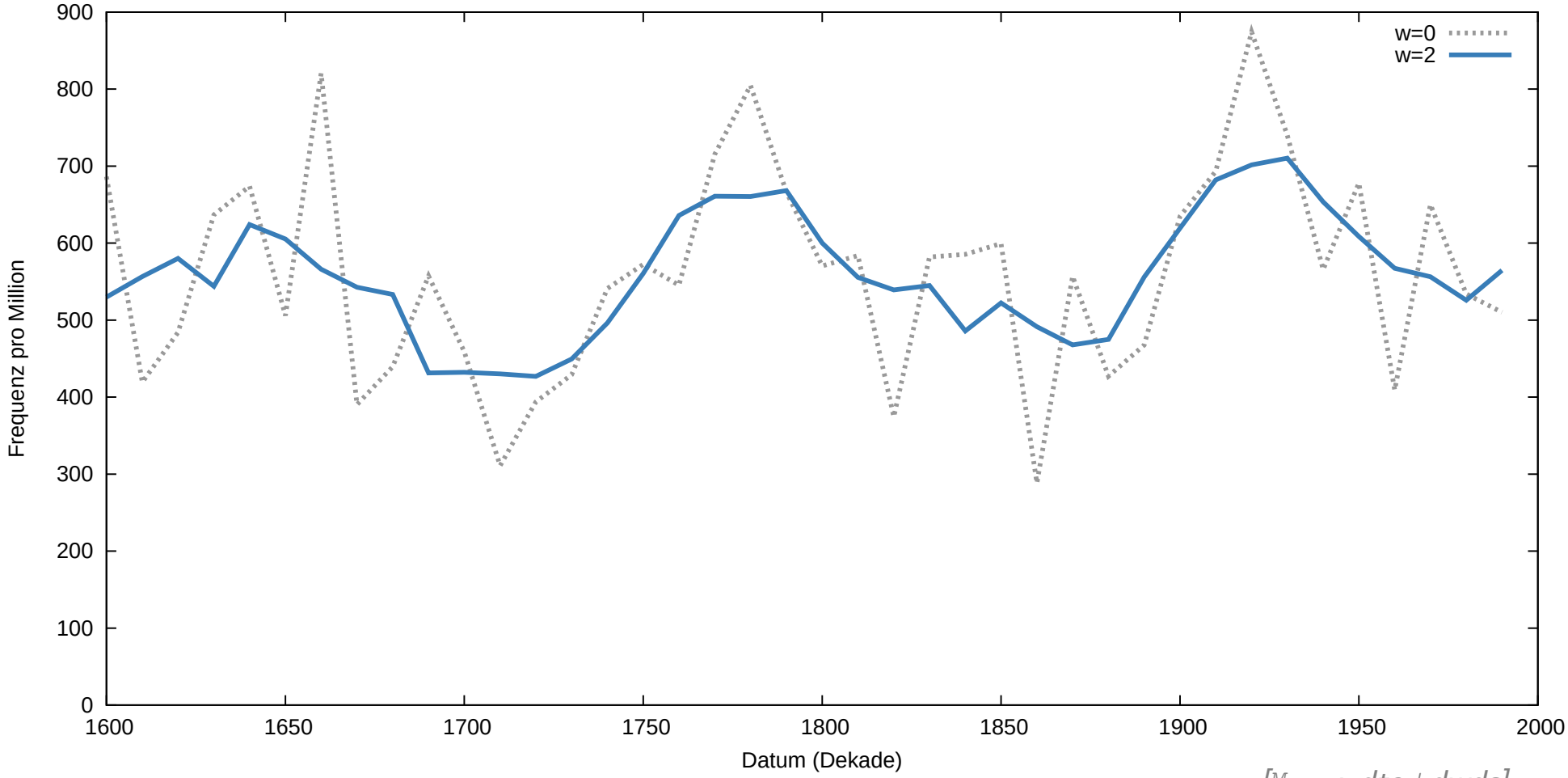
- **Problem:** gemessene Frequenzen schwanken lokal ("zick-zack")
- **Ursache:** (zu) kleine Intervalle, spärliche Daten, "Rauschen"
- **Lösung:** *gleitender Mittelwert* über w Vorgänger/Nachfolger (w=1)



[Mann: dta+dwds]

Glättung: lokale Schwankungen

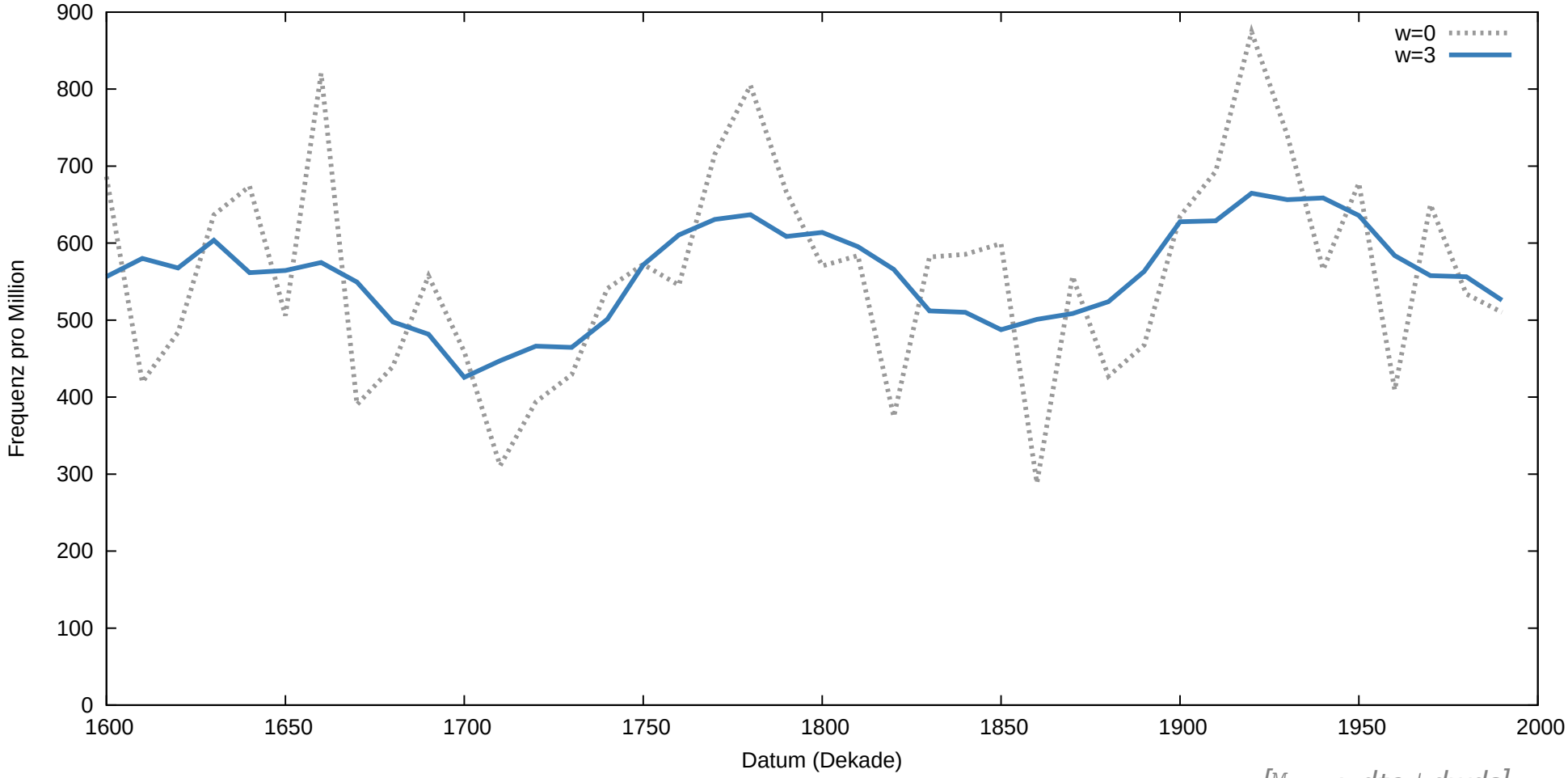
- **Problem:** gemessene Frequenzen schwanken lokal ("zick-zack")
- **Ursache:** (zu) kleine Intervalle, spärliche Daten, "Rauschen"
- **Lösung:** *gleitender Mittelwert* über w Vorgänger/Nachfolger (w=2)



[Mann: dta+dwds]

Glättung: lokale Schwankungen

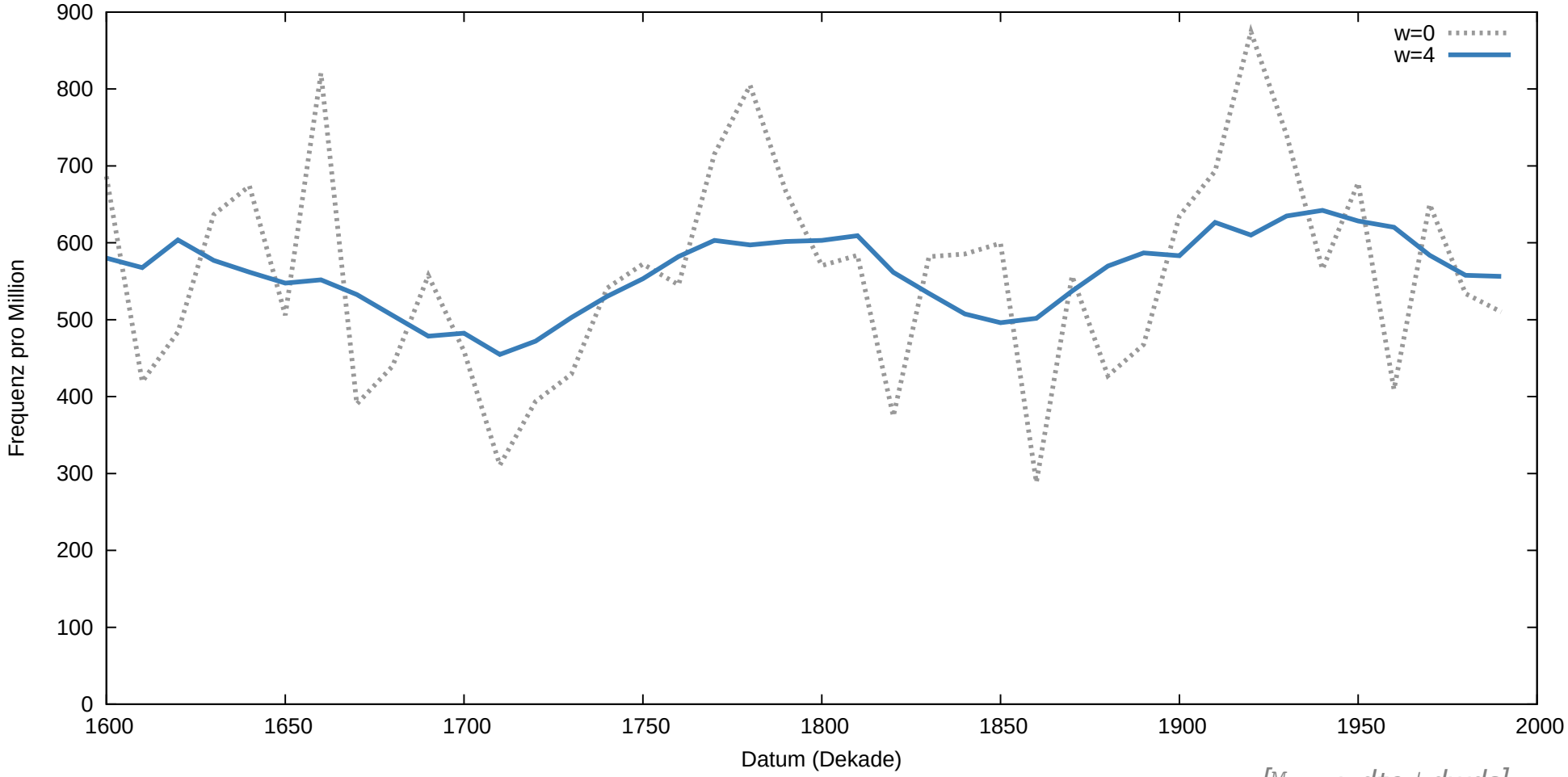
- **Problem:** gemessene Frequenzen schwanken lokal ("zick-zack")
- **Ursache:** (zu) kleine Intervalle, spärliche Daten, "Rauschen"
- **Lösung:** *gleitender Mittelwert* über w Vorgänger/Nachfolger (w=3: Default)



[Mann: dta+dwds]

Glättung: lokale Schwankungen

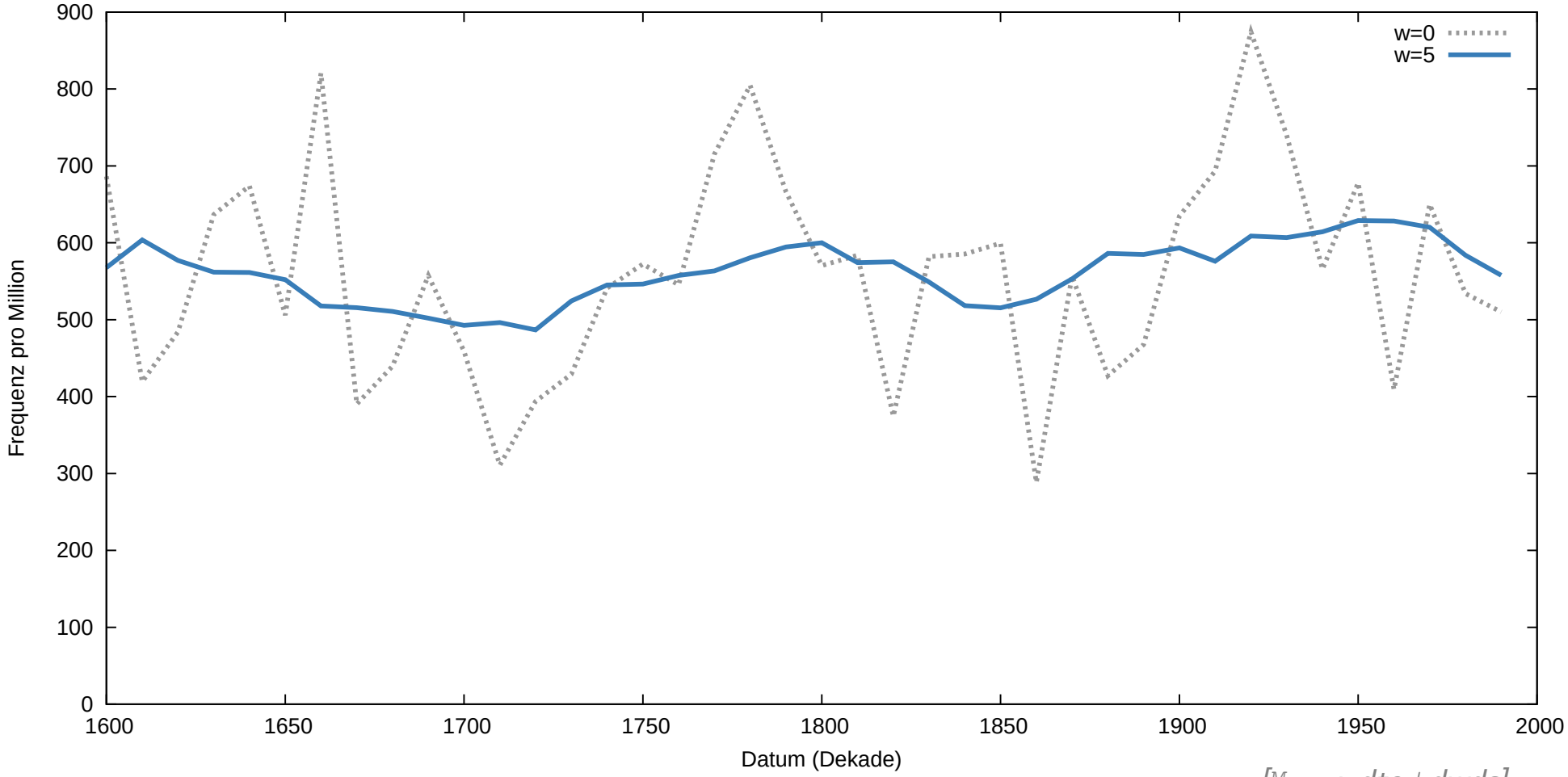
- **Problem:** gemessene Frequenzen schwanken lokal ("zick-zack")
- **Ursache:** (zu) kleine Intervalle, spärliche Daten, "Rauschen"
- **Lösung:** *gleitender Mittelwert* über w Vorgänger/Nachfolger (w=4)



[Mann: dta+dwds]

Glättung: lokale Schwankungen

- **Problem:** gemessene Frequenzen schwanken lokal ("zick-zack")
- **Ursache:** (zu) kleine Intervalle, spärliche Daten, "Rauschen"
- **Lösung:** *gleitender Mittelwert* über w Vorgänger/Nachfolger (w=5)



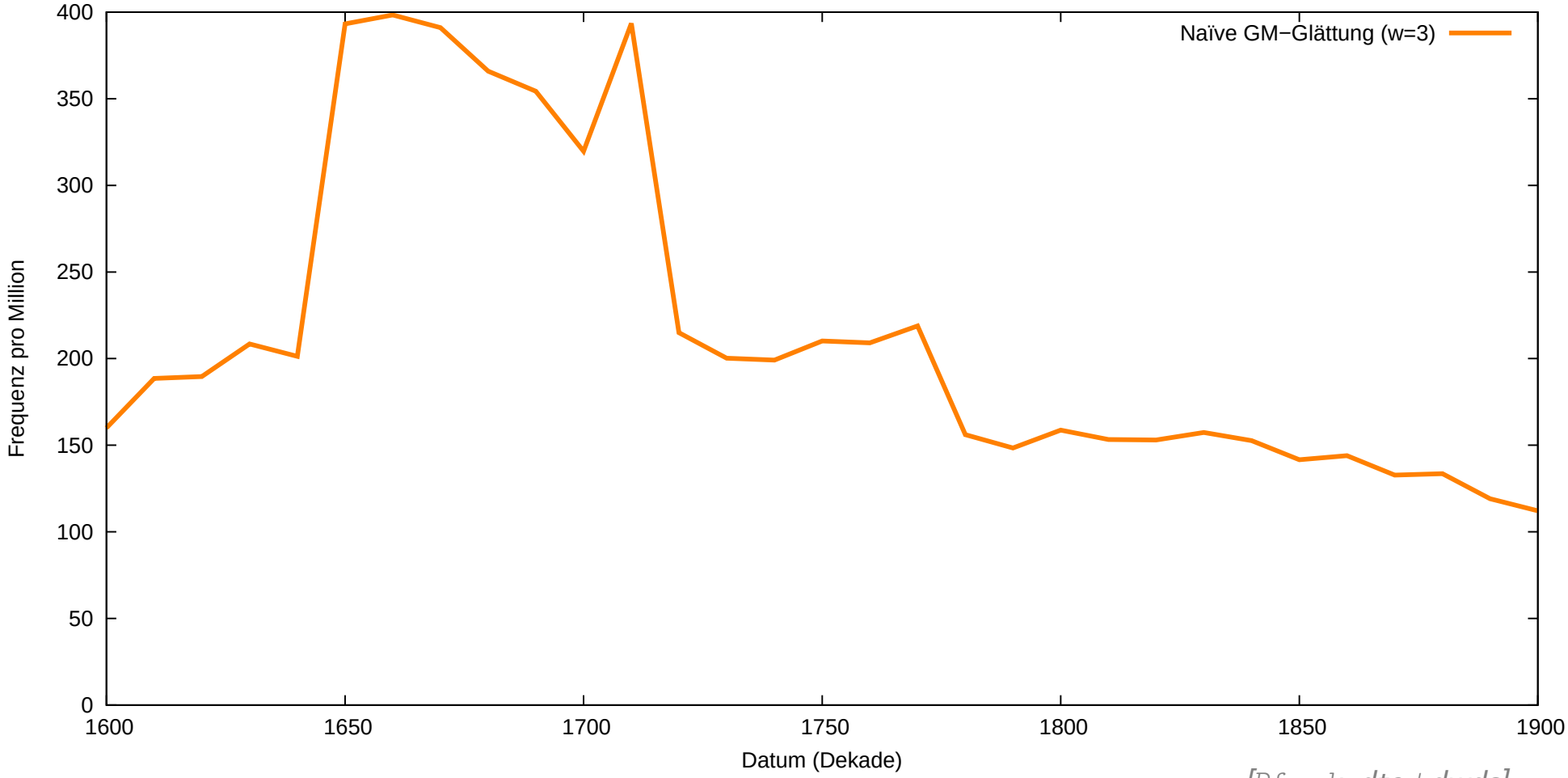
[Mann: dta+dwds]

Ausreißererkennung (“Pruning”)

- **Problem:** “Kurvenbuckel” $\sim$ Treppenfunktion *(Glättungsartefakte)*

Ausreißererkennung (“Pruning”)

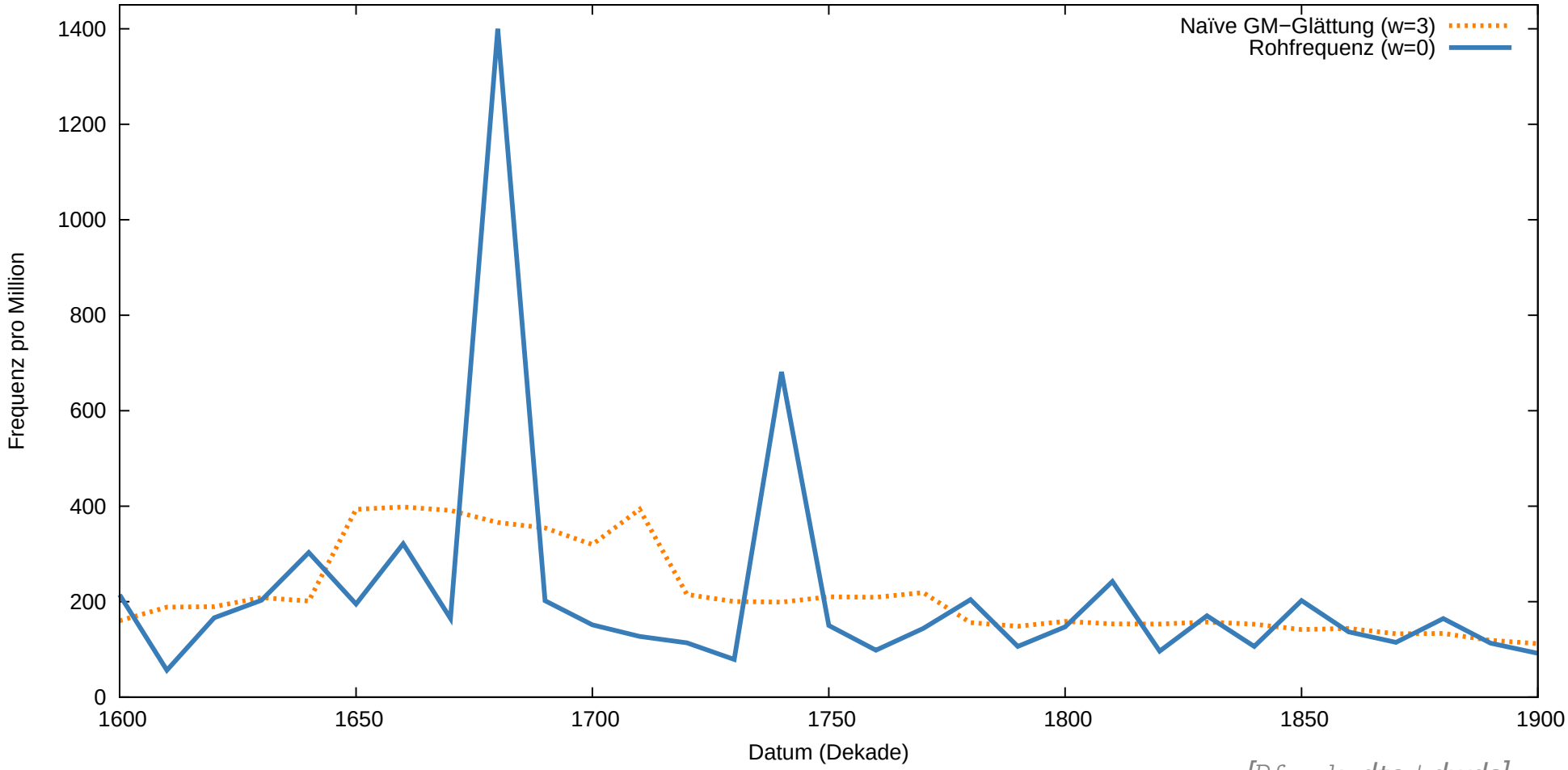
- **Problem:** “Kurvenbuckel” $\sim$ Treppenfunktion *(Glättungsartefakte)*



[Pferd: dta+dwds]

Ausreißererkennung (“Pruning”)

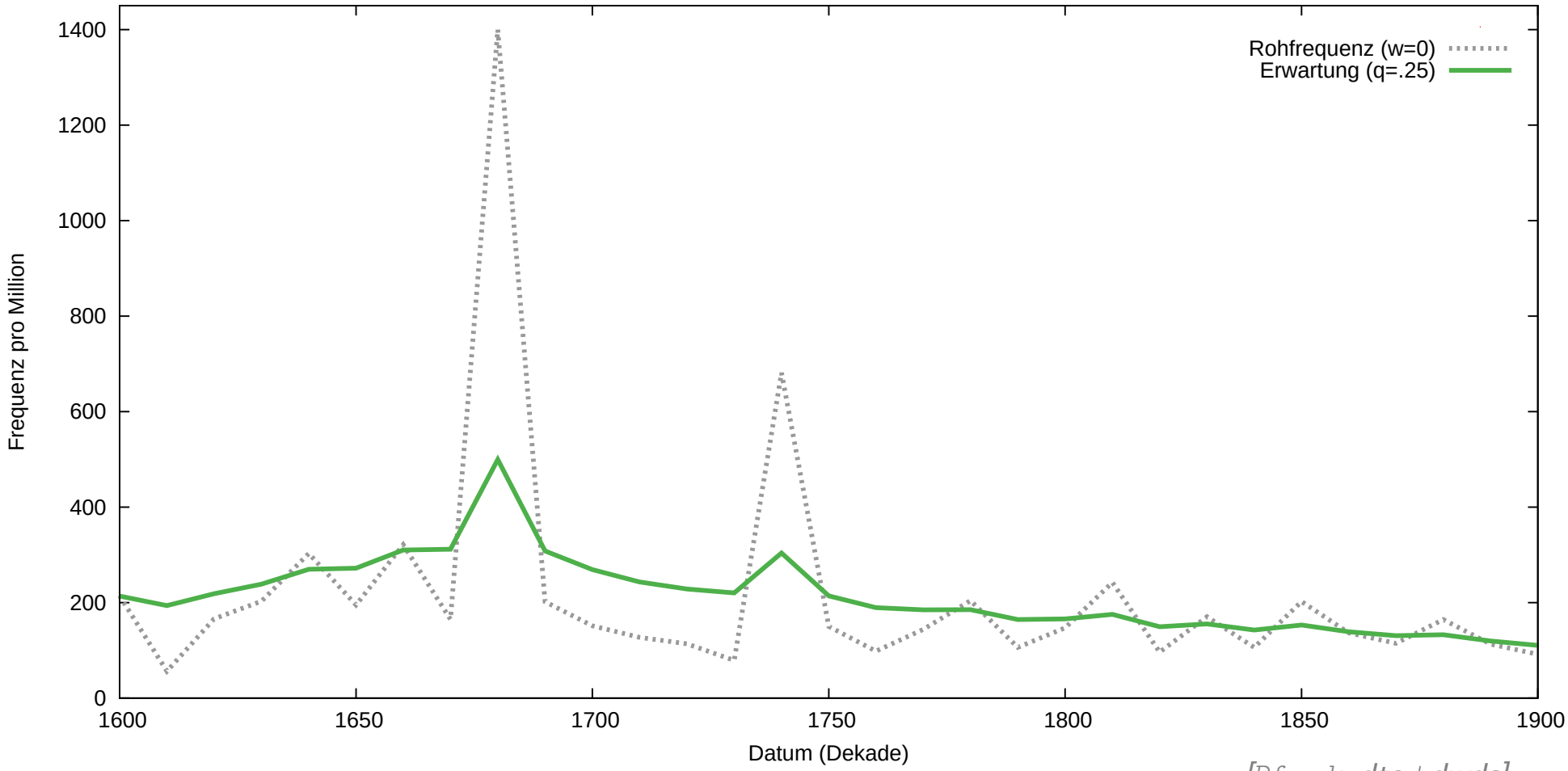
- **Problem:** “Kurvenbuckel” ∼ Treppenfunktion *(Glättungsartefakte)*
- **Ursache:** Korpus ist *nicht repräsentativ* *(Stichprobenartefakte)*
  - J. C. Pinter von der Au, *Pferd-Schatz*, 1688.
  - J. Trichter, *Ritter-Exercitien-Lexicon*, 1742.



[Pferd: dta+dwds]

Ausreißererkennung (“Pruning”)

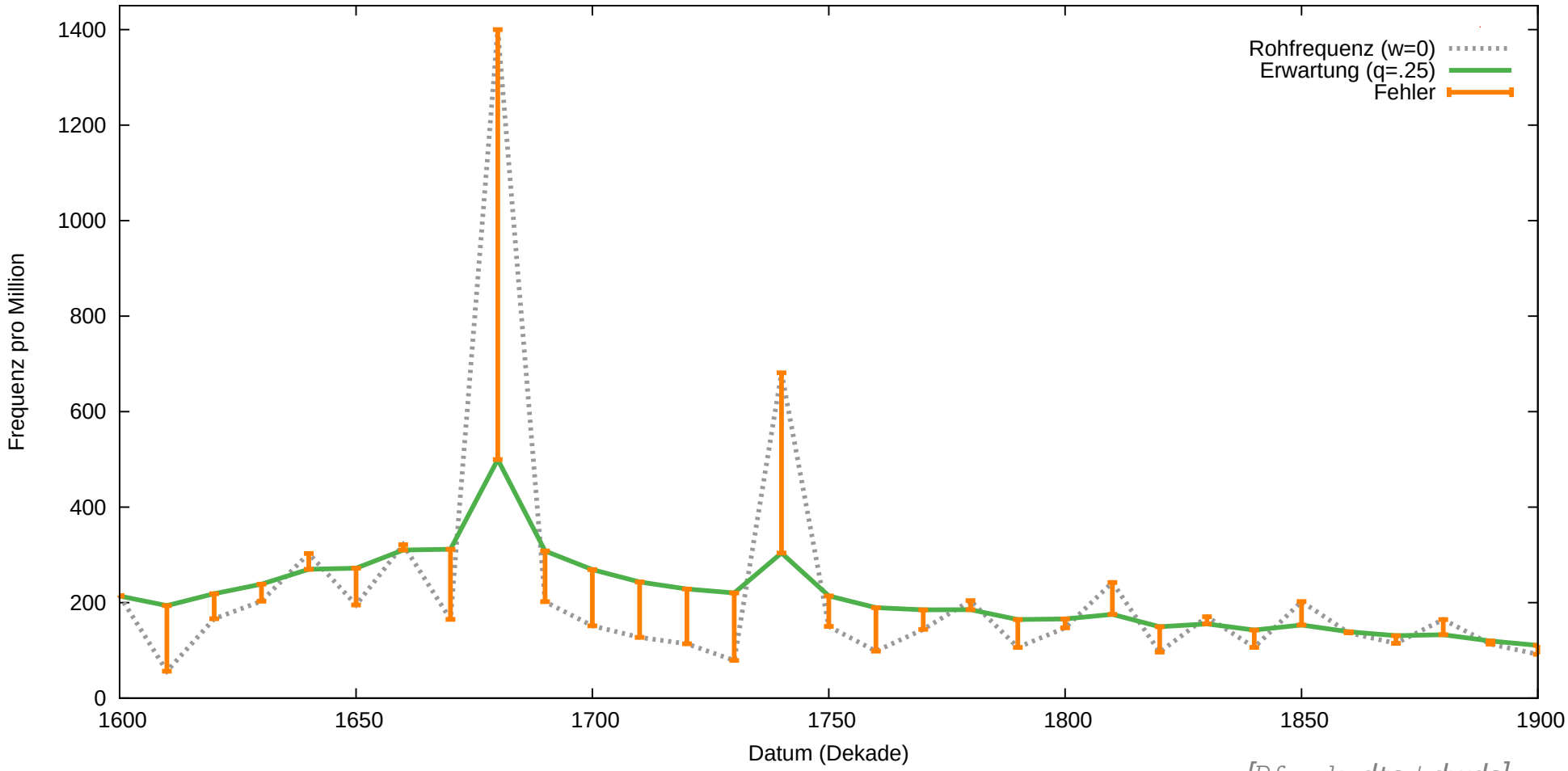
- **Problem:** Korpusartefakte ⇝ “Kurvenbuckel”
- **Lösung:** Ausreißererkennung & -entfernung
- **Schritt 1:** “Erwartungsfunktion” berechnen (doppelter exp. Filter)



[Pferd: dta+dwds]

Ausreißererkennung (“Pruning”)

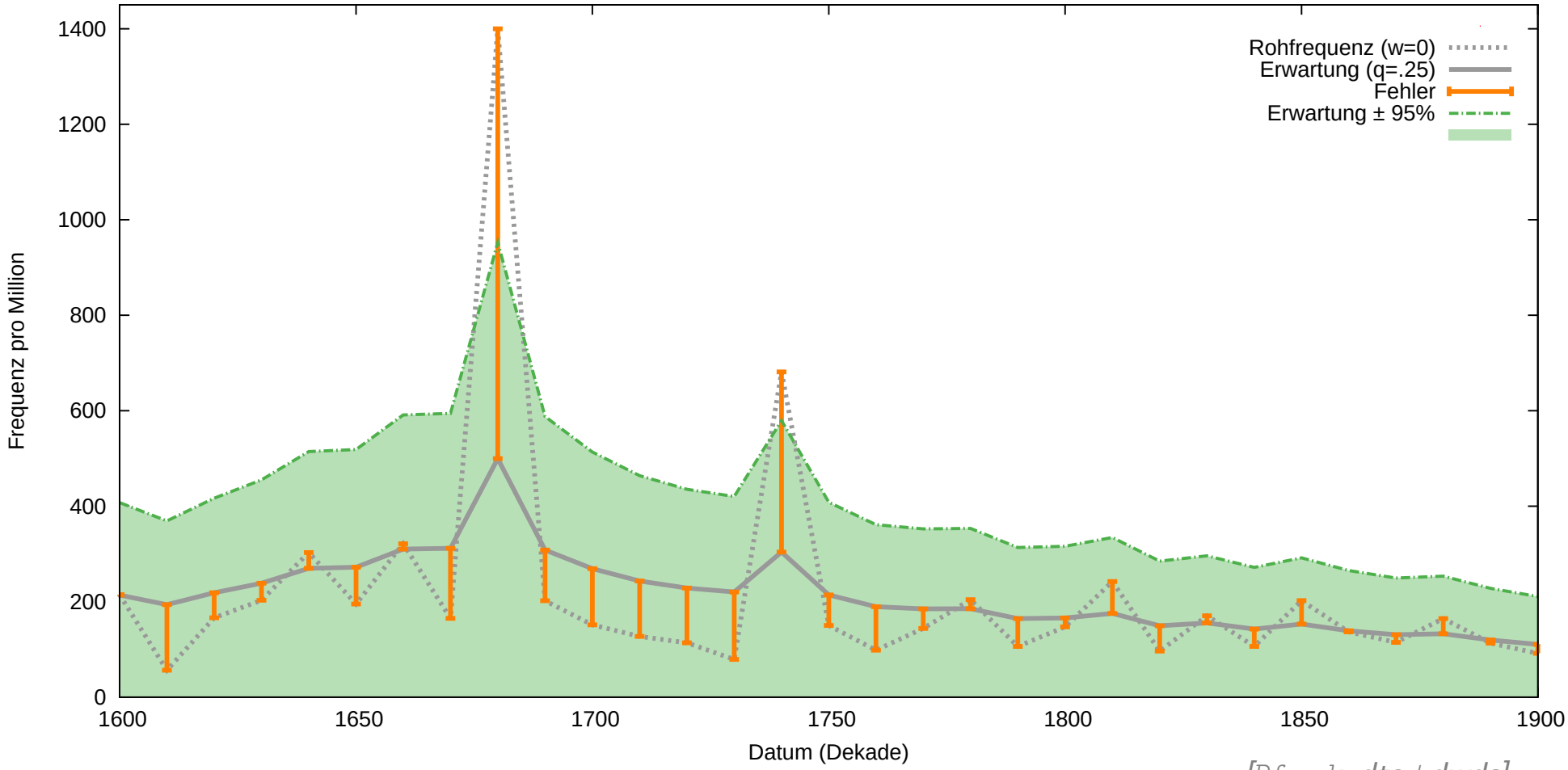
- **Problem:** Korpusartefakte $\rightsquigarrow$ “Kurvenbuckel”
- **Schritt 1:** “Erwartungsfunktion” berechnen (doppelter exp. Filter)
- **Schritt 2:** erzeuge *Fehlerverteilung* bzgl. Erwartung $\sim N(\mu, \sigma)$



[Pferd: dta+dwds]

Ausreißererkennung ("Pruning")

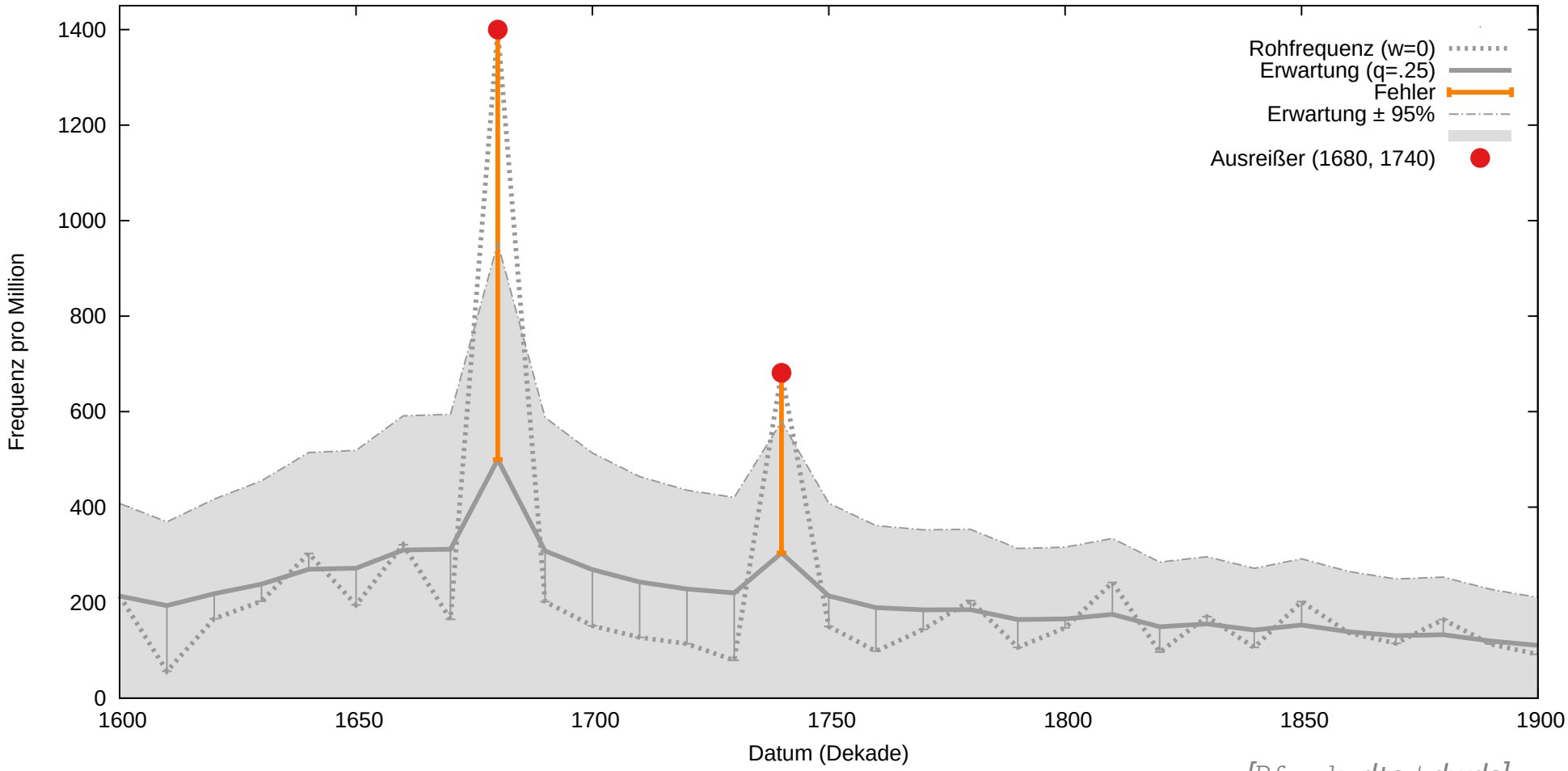
- **Problem:** Korpusartefakte $\rightsquigarrow$ "Kurvenbuckel"
- **Schritt 2:** erzeuge *Fehlerverteilung* bzgl. Erwartung $\sim N(\mu, \sigma)$
- **Schritt 3:** projeziere *Konfidenzintervall* auf Datenraum $p = 95\%$



[Pferd: dta+dwds]

Ausreißererkennung (“Pruning”)

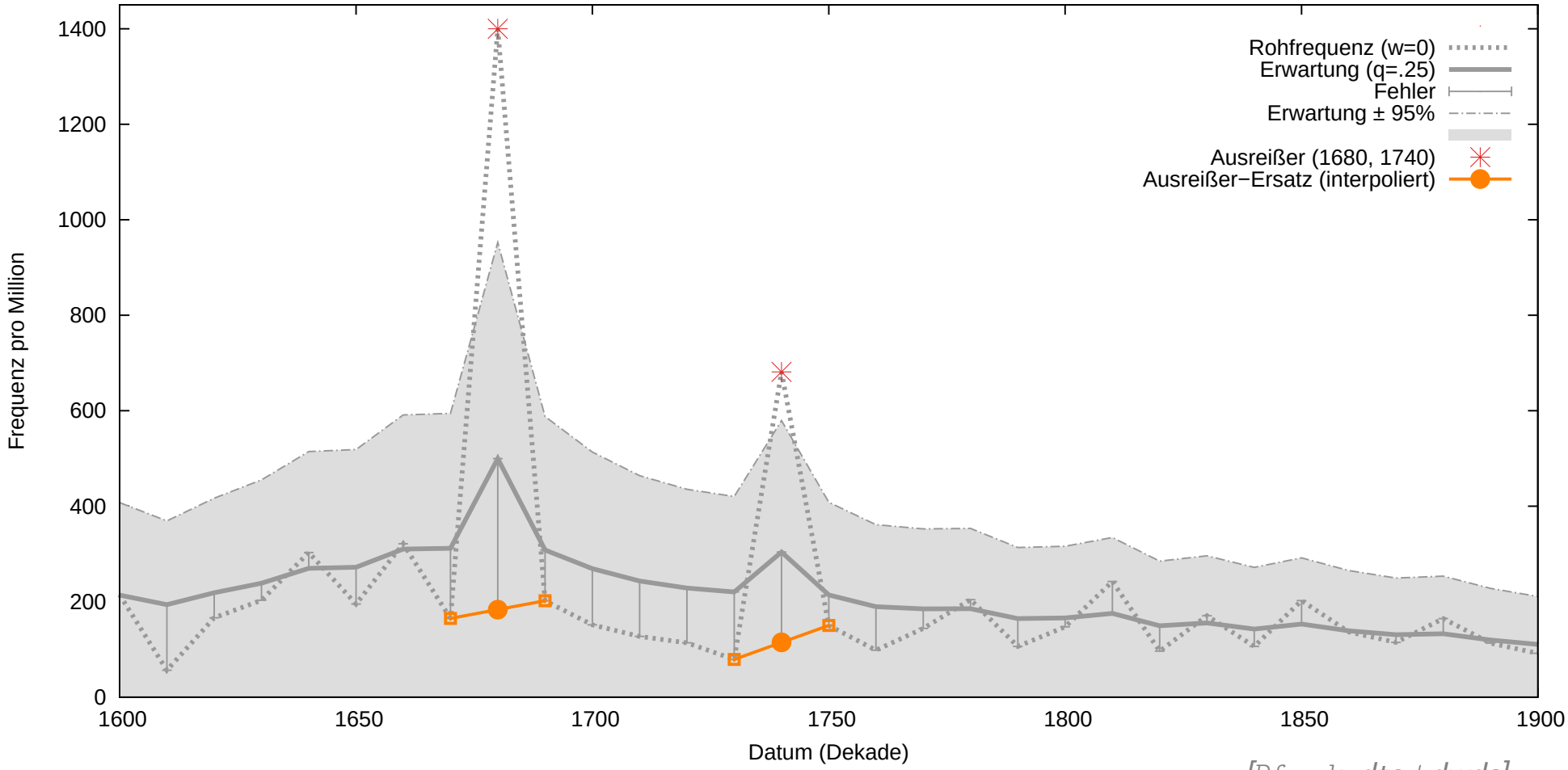
- **Problem:** Korpusartefakte $\rightsquigarrow$ “Kurvenbuckel”
- **Schritt 3:** projeziere *Konfidenzintervall* auf Datenraum $p = 95\%$
- **Schritt 4:** “Ausreißer” = Punkte jenseits des projezierten Konfidenzintervalls



[Pferd: dta+dwds]

Ausreißererkennung (“Pruning”)

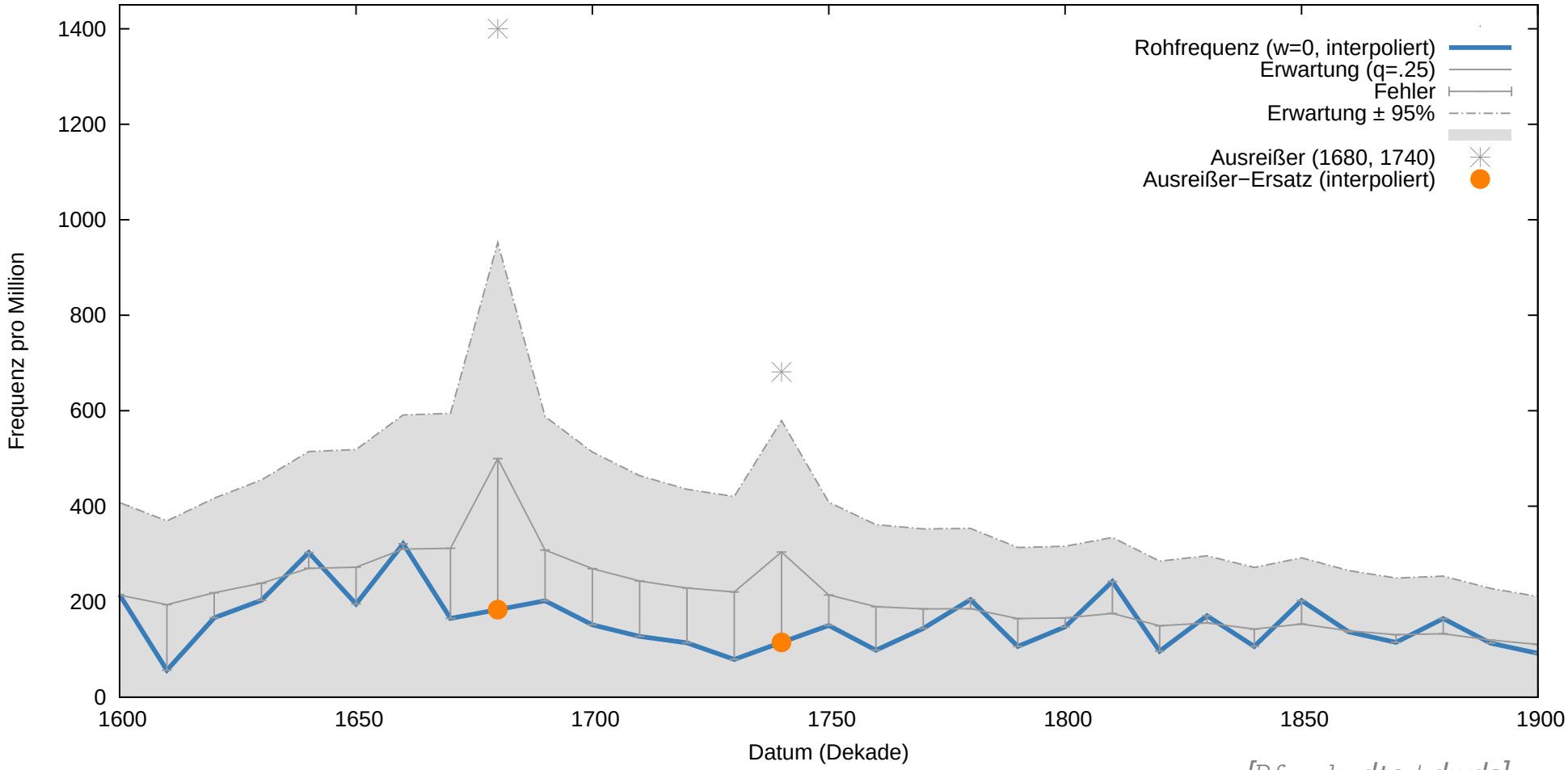
- **Problem:** Korpusartefakte ⇝ “Kurvenbuckel”
- **Schritt 4:** “Ausreißer” = Punkte jenseits des projezierten Konfidenzintervalls
- **Schritt 5:** ersetze Ausreißer (lineare Interpolierung mit “guten” Nachbarn)



[Pferd: dta+dwds]

Ausreißererkennung (“Pruning”)

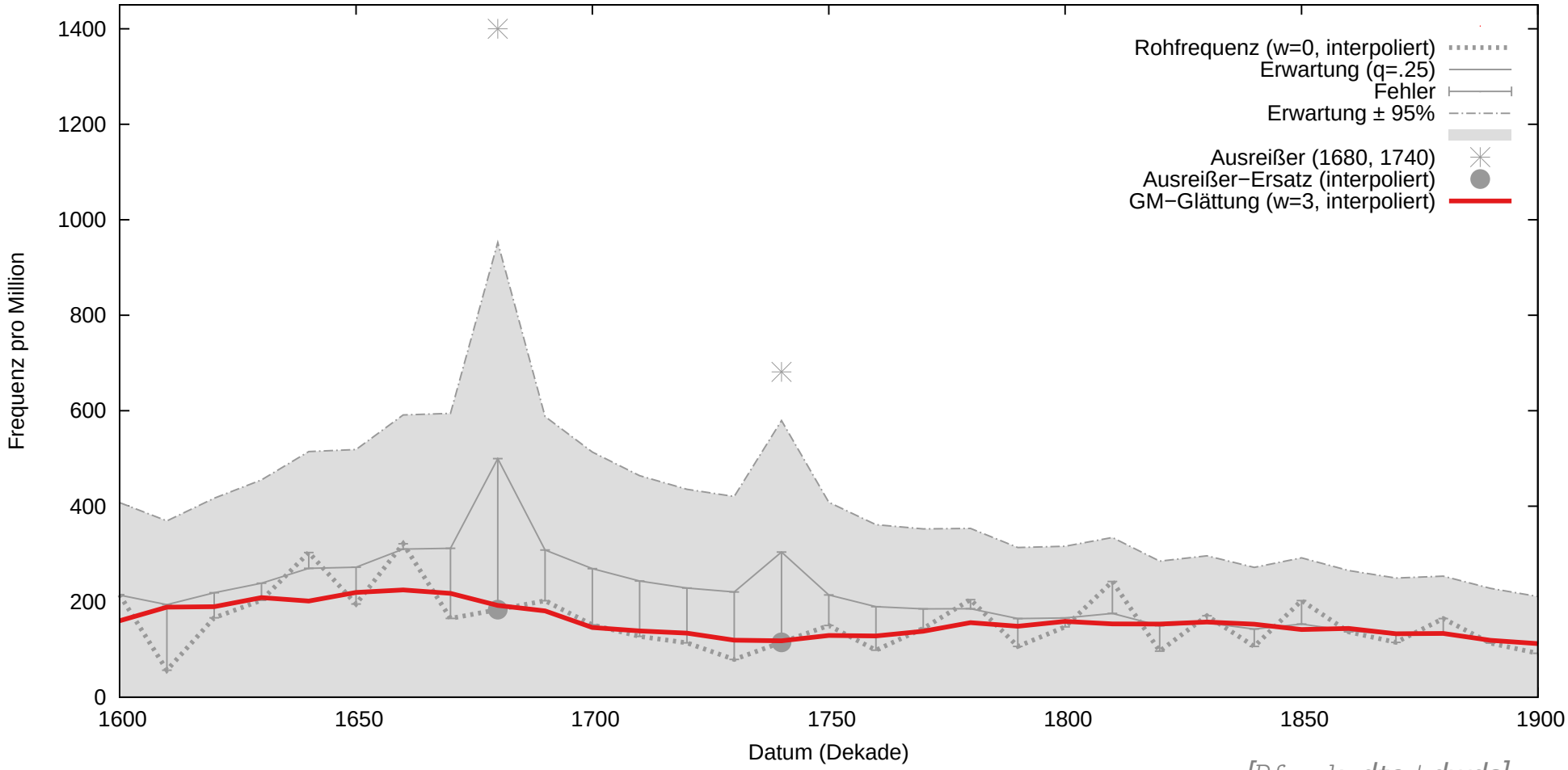
- **Problem:** Korpusartefakte ⇝ “Kurvenbuckel”
- **Schritt 4:** “Ausreißer” = Punkte jenseits des projezierten Konfidenzintervalls
- **Schritt 5:** ersetze Ausreißer (lineare Interpolierung mit “guten” Nachbarn)



[Pferd: dta+dwds]

Ausreißererkennung (“Pruning”)

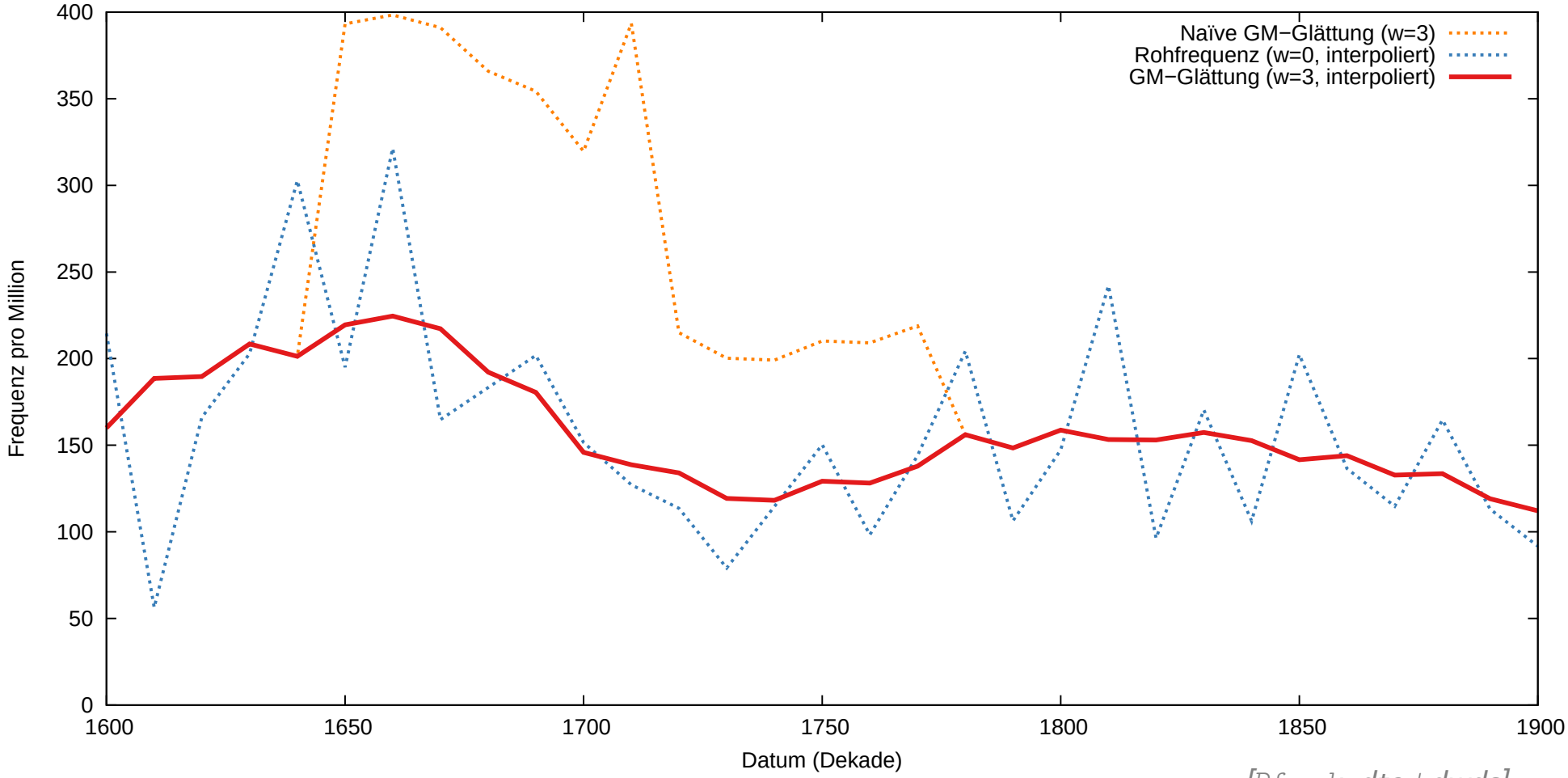
- **Problem:** Korpusartefakte ⇝ “Kurvenbuckel”
- **Schritt 5:** ersetze Ausreißer (lineare Interpolierung mit “guten” Nachbarn)
- **Schritt 6:** gleitender Mittelwert bzgl. “reparierten” Daten



[Pferd: dta+dwds]

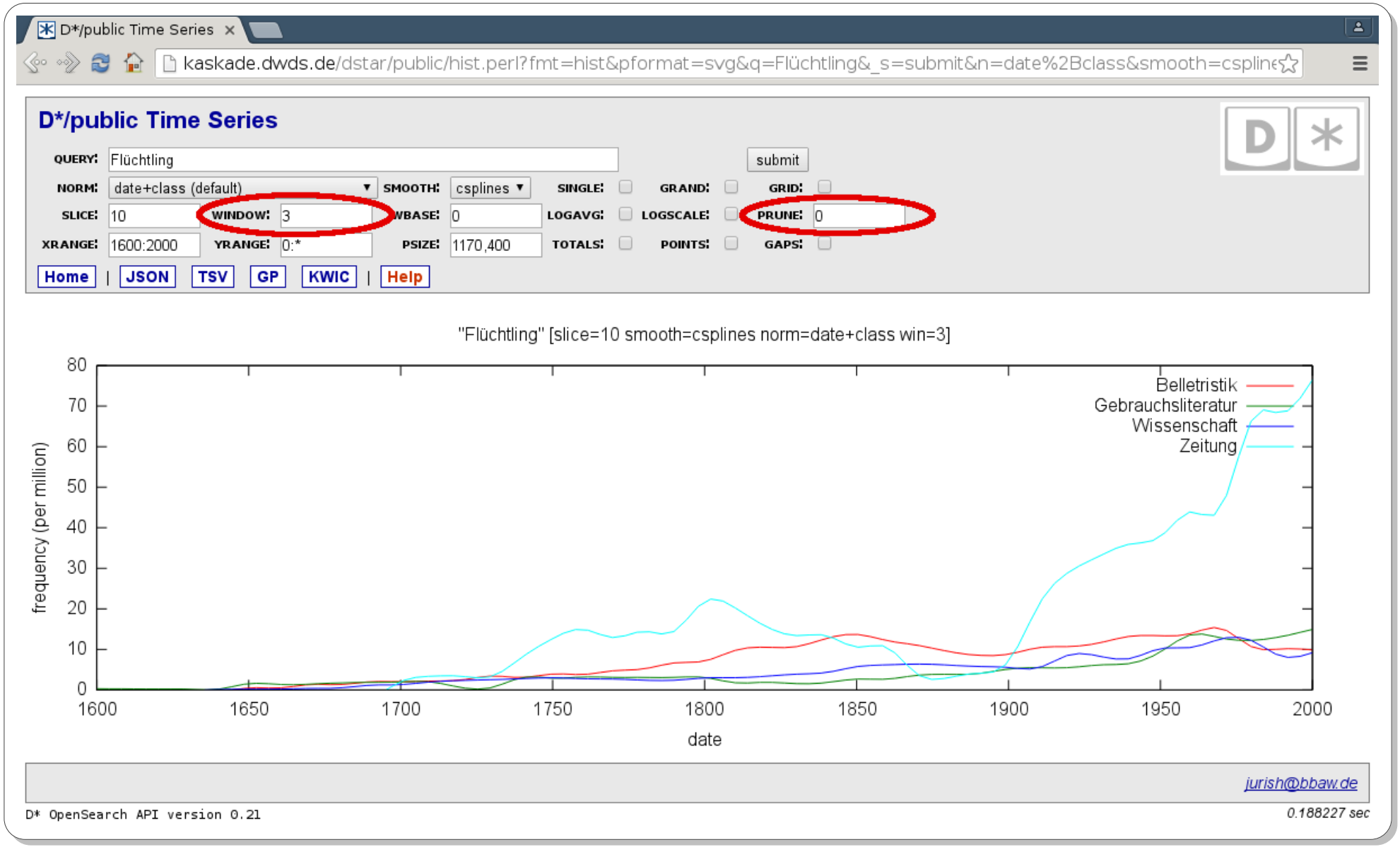
Ausreißererkennung (“Pruning”)

- **Problem:** Korpusartefakte ⇝ “Kurvenbuckel”
- **Schritt 6:** gleitender Mittelwert bzgl. “reparierten” Daten
- *. . . fertig!*



[Pferd: dta+dwds]

Integrierung: DDC/D* WWW API



Integrierung: `zwei.dwds.de`

DWDS [Betaversion]

zwei.dwds.de/wb/Flüchtling

D W D S — Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. [BETAVERSION] — Anmelden

Flüchtling

Flüchtling – Substantiv
Maskulinum, -s, -e
Aussprache:
Wortzerlegung: ↗flüchten ↗-ling
Worttrennung: Flücht-ling (computergeneriert)

Bedeutung | Etymologie | Thesaurus | Verwendungsbeispiele | Typische Verbindungen | Wortbildung

Bedeutung

eWDG, 1967

jmd., der vor jmdm., etw. geflüchtet ist und dabei alles verloren hat

Beispiele:

ein heimatloser, politischer **Flüchtling**

ein Strom, eine Kolonne von **Flüchtlingen**

ein Lager für **Flüchtlinge**

In Frankreich lebten die italienischen politischen **Flüchtlinge** [Lamprecht, *Dt. Geschichte*, 373]

es sind neue Transporte mit Umsiedlern, **Flüchtlingen**, eingetroffen, und die Behörde wird mit der Einbürgerung nicht fertig [Baierl, *Flinz*, 1]

Von überallher strömten **Flüchtlinge** ein [Fussenegger, *Antlitz*, 197]

Startseite / Wortinformation zu „Flüchtling"

Worthäufigkeit

selten — häufig

Frequenz / Mio Tokens

75 50 25 0 1700 1800 1900

Ältere Wörterbücher

- Grimmsches Wörterbuch (¹DWB) (1)
- Wörterbuch der deutschen Gegenwartssprache (WDG) (1)

Korpustreffer

Referenzkorpora

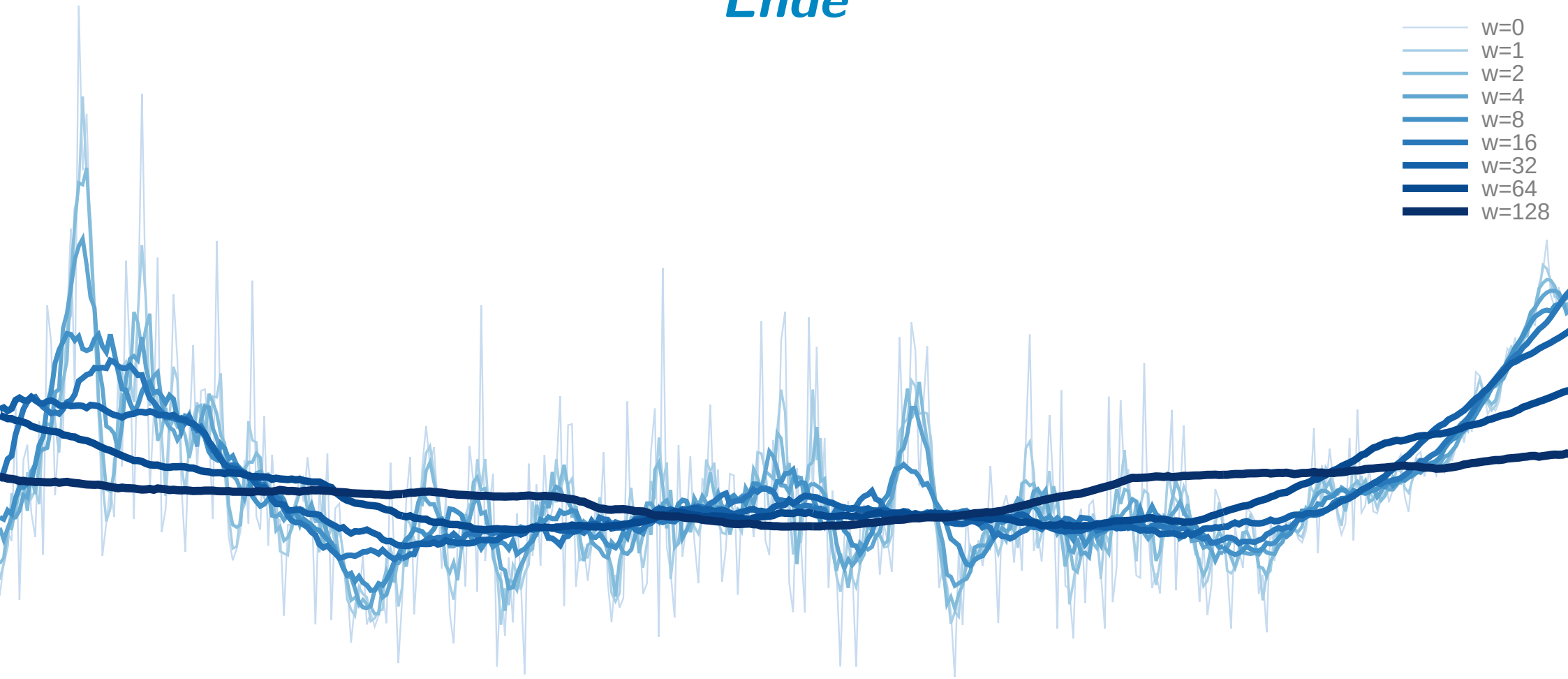
- DWDS-Kernkorpus (2523)
- DWDS-Kernkorpus 21 (223)
- Deutsches Textarchiv (801)

Zusammenfassung

Wortverläufe

- Zeitabhängige Häufigkeitsverteilungen
- Indikator für diachrone lexikalische Prozesse
- Intuitive graphische Darstellung

Besonderheiten

- Schreibweisentoleranz
- Flexible Such- & Filteroptionen
- Direkte Verknüpfung mit Korpustreffern
- Kurvenglättung & Ausreißererkennung
- Externes API $\rightsquigarrow$ Weiternutzung

“Ende”