Statistical Methods for Dating Collections of Historical Documents

Michael Gervers

University of Toronto

—

- Problem – Statistical methodologies for dating documents and texts.

- Motivation – Historians want to date source documents accurately.

The Data

- A total of 3353 documents which have all been accurately dated by historians.
- These documents are in digitized format.
- The 3353 documents were divided into a training set, validation set and test set.

- The training documents “teach” or “train” our dating algorithm.
- The validation set is used for estimating certain parameters.
- The test set is used to measure accuracy.

ID: 00640214
Document date: 1237

Haec est finalis concordia facta in curia domini regis apud Westmonasterium a die S Johannis Baptistae in !xv! dies anno regni regis Henrici filii regis Johannis !xxi! coram Roberto de Lexinton Willelmo de Eboraco Ada filio Willelmi Willelmo de Culewurth justitiariis et aliis domini regis fidelibus tunc ibi praesentibus inter Johannem Baioc quaerentem et Robertum Sarum episcopum et capitulum .....

- The concept of shingles
- A *shingle* is a consecutive sequence of words (Broder, 1998).
- Example:

$$\mathcal{D} = (\text{a rose is a rose is a rose})$$

then the set of its *k-shingles* (say, $k = 2$) is:

$$S_2(\mathcal{D}) = \{\{\text{a rose}\}, \{\text{rose is}\}, \{\text{is a}\}, \{\text{a rose}\}, \{\text{rose is}\}, \{\text{is a}\}, \{\text{a rose}\}\}$$

The idea behind the maximum prevalence method

To date an undated document $\mathcal{D}$:

1) Construct the set $S(\mathcal{D})$ for a fixed shingle order.

The idea behind the maximum prevalence method

To date an undated document $\mathcal{D}$:

1) Construct the set $S(\mathcal{D})$ for a fixed shingle order.

2) For each shingle in the set $S(\mathcal{D})$, estimate the probability of its occurrence as a function of time.

The idea behind the maximum prevalence method

To date an undated document $\mathcal{D}$:

1) Construct the set $S(\mathcal{D})$ for a fixed shingle order.

2) For each shingle in the set $S(\mathcal{D})$, estimate the probability of its occurrence as a function of time.

3) Combine the probability of occurrence of the shingles together.

The idea behind the maximum prevalence method

To date an undated document $\mathcal{D}$:

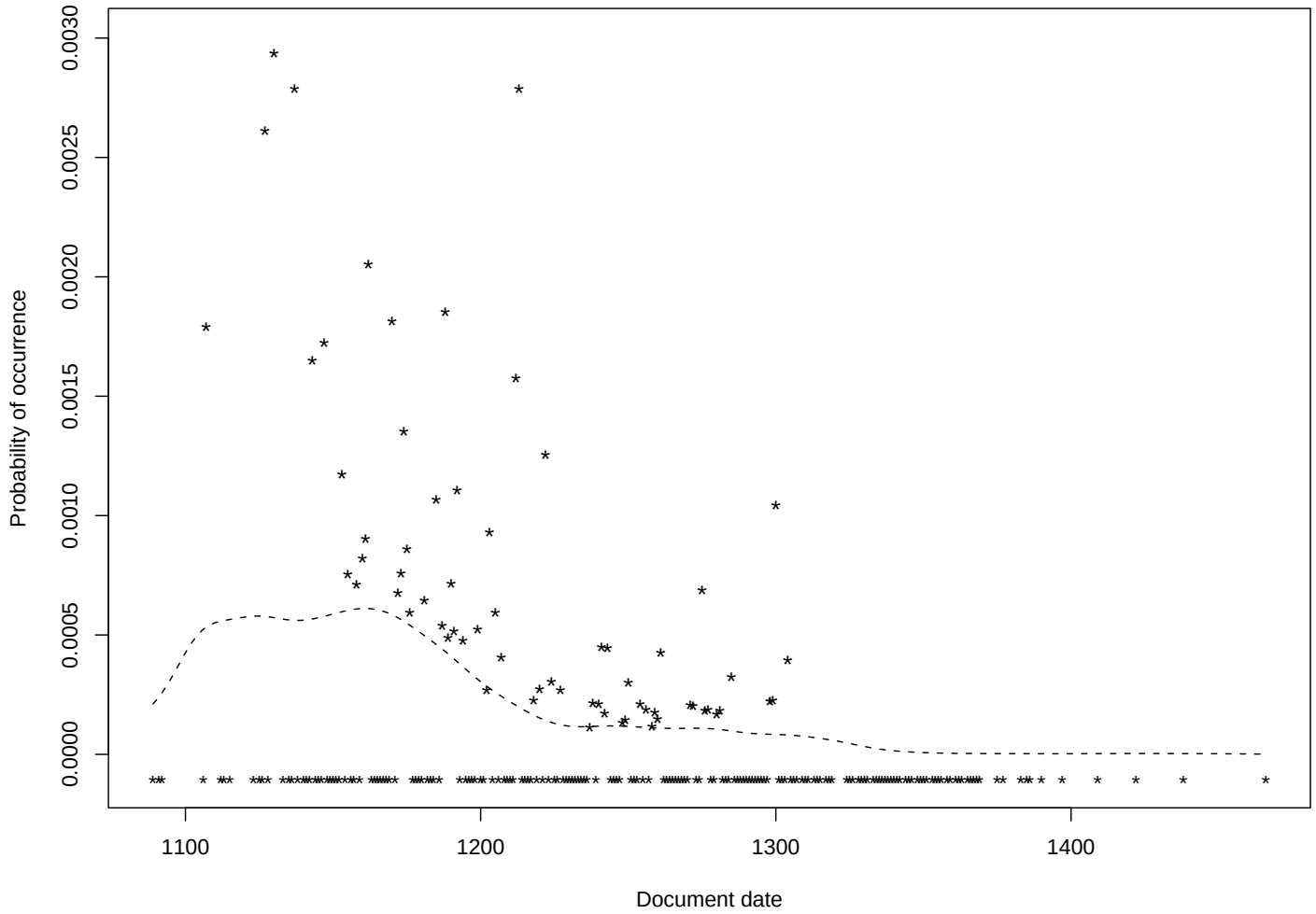
1) Construct the set $S(\mathcal{D})$ for a fixed shingle order.

2) For each shingle in the set $S(\mathcal{D})$, estimate the probability of its occurrence as a function of time.

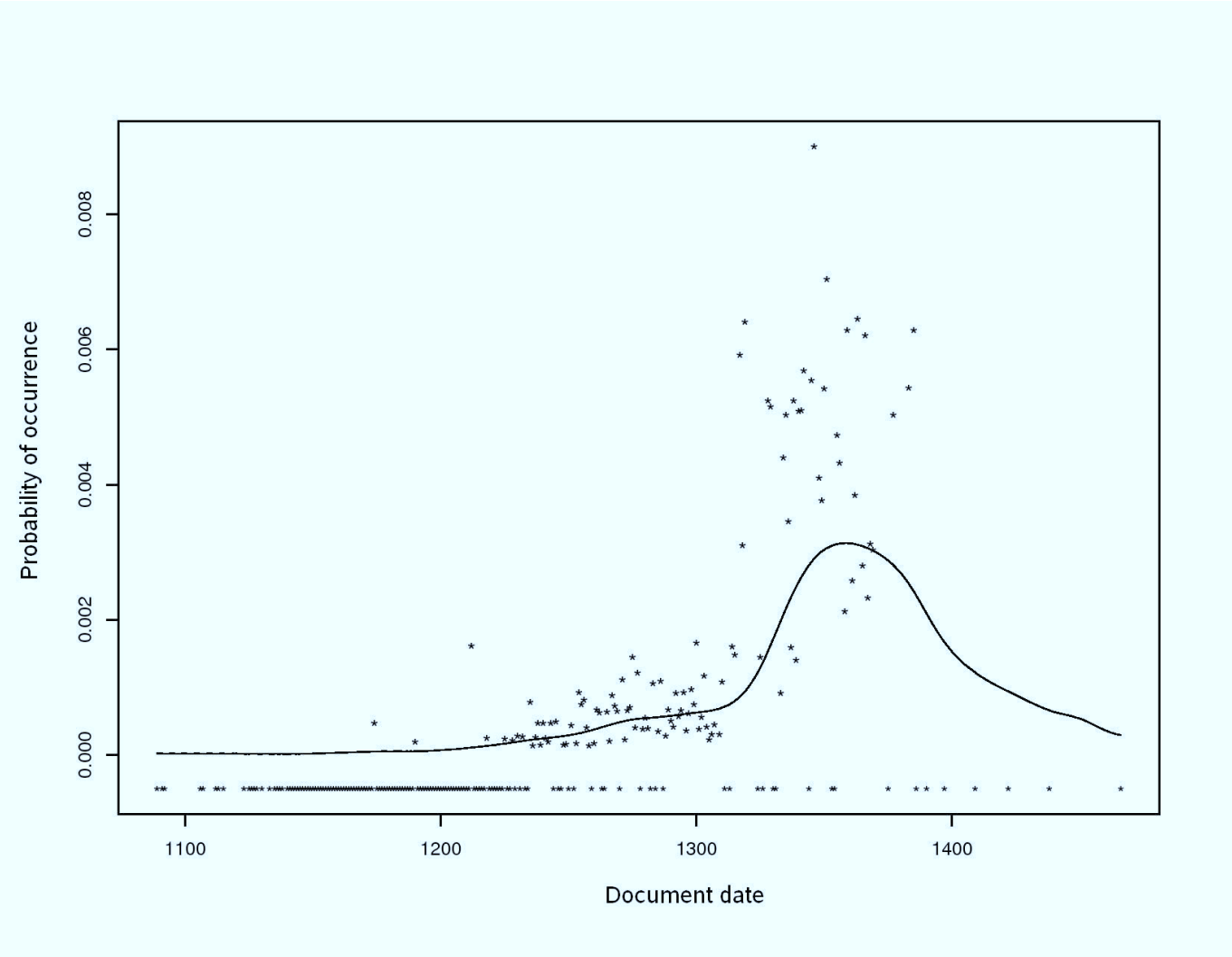
3) Combine the probability of occurrence of the shingles together.

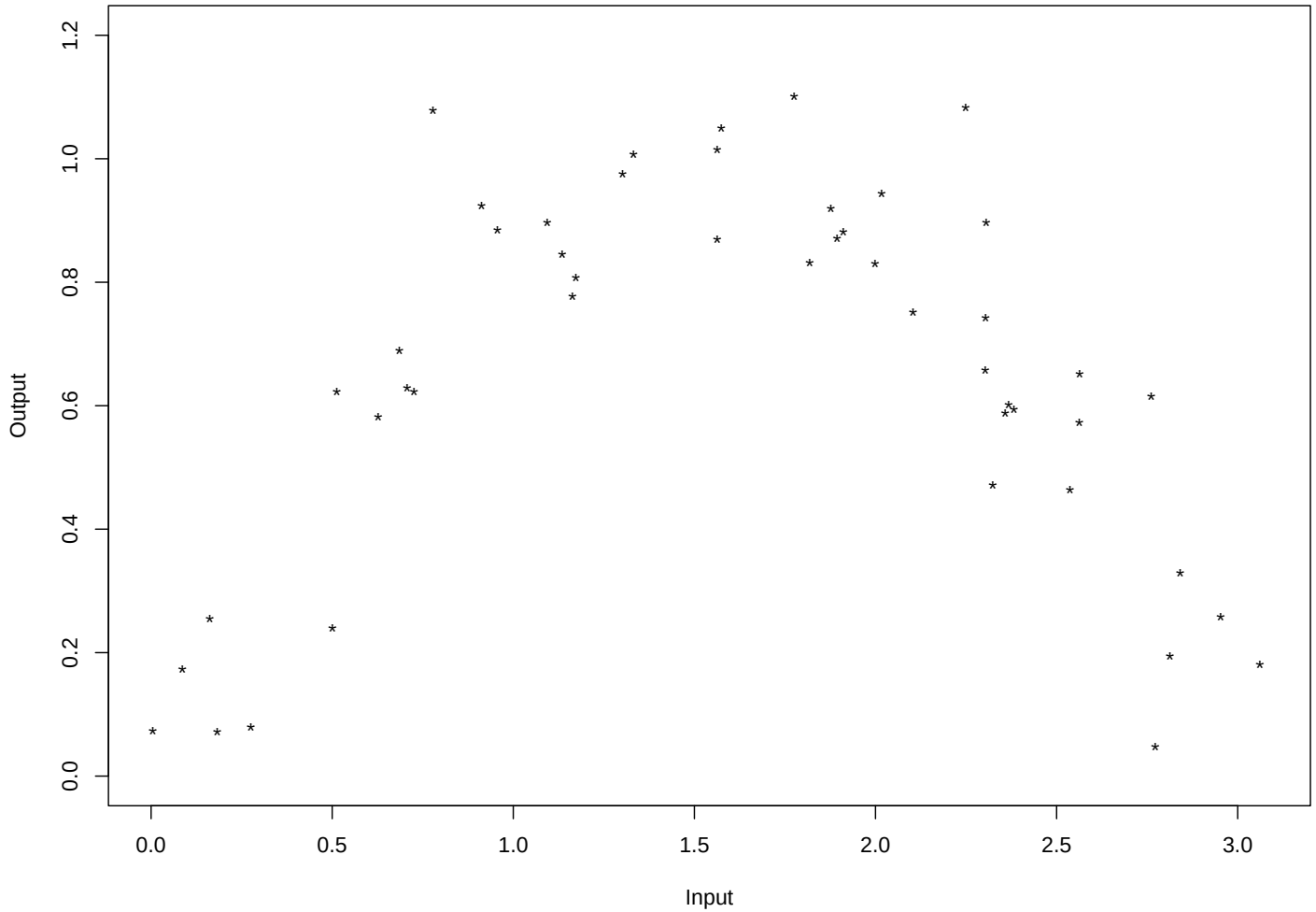
4) The value where the peak of the resulting function occurs is taken to be the date estimate of document $\mathcal{D}$.

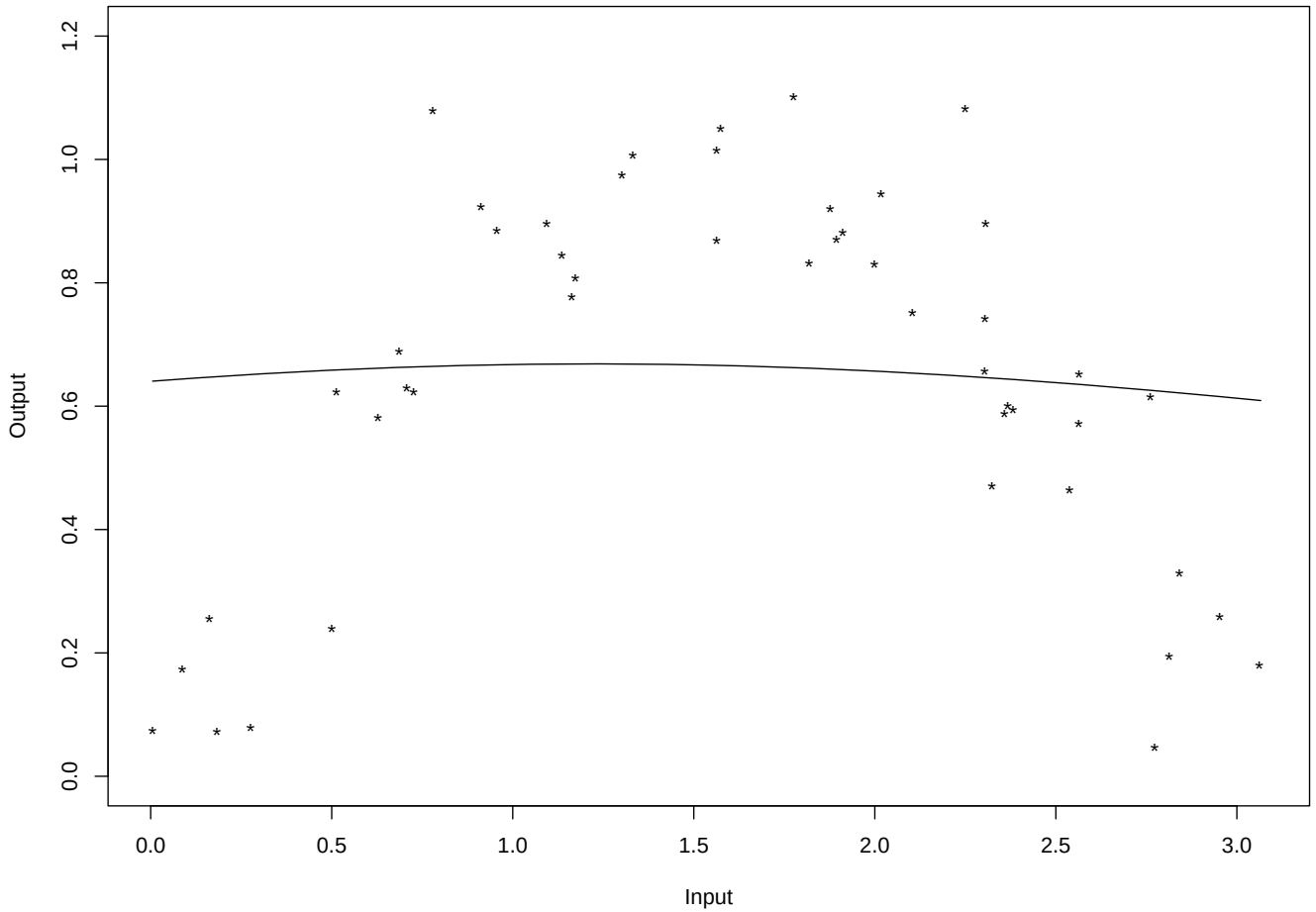
The probability of occurrence of the shingle *ibidem Deo seruientibus* as a function of time

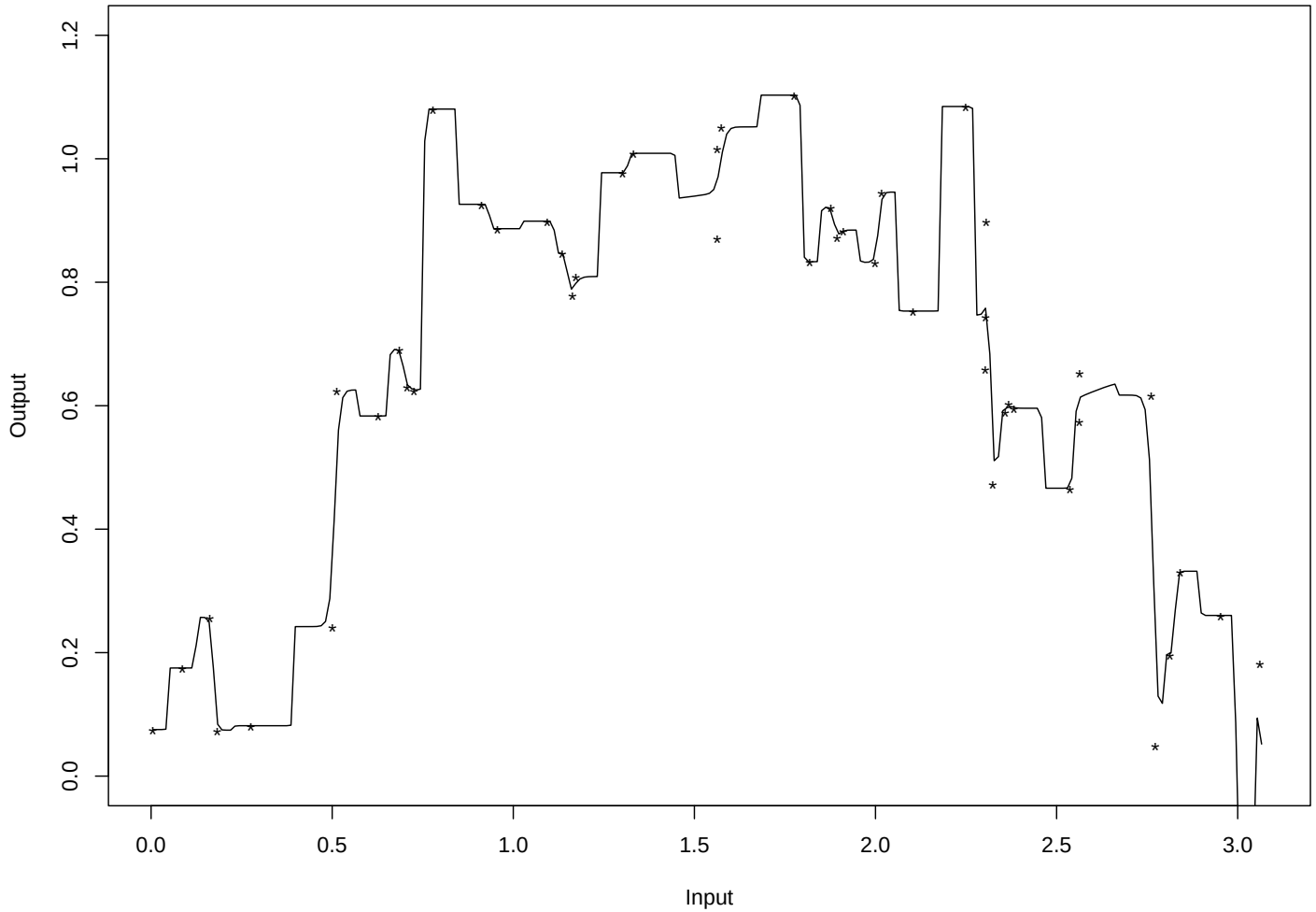


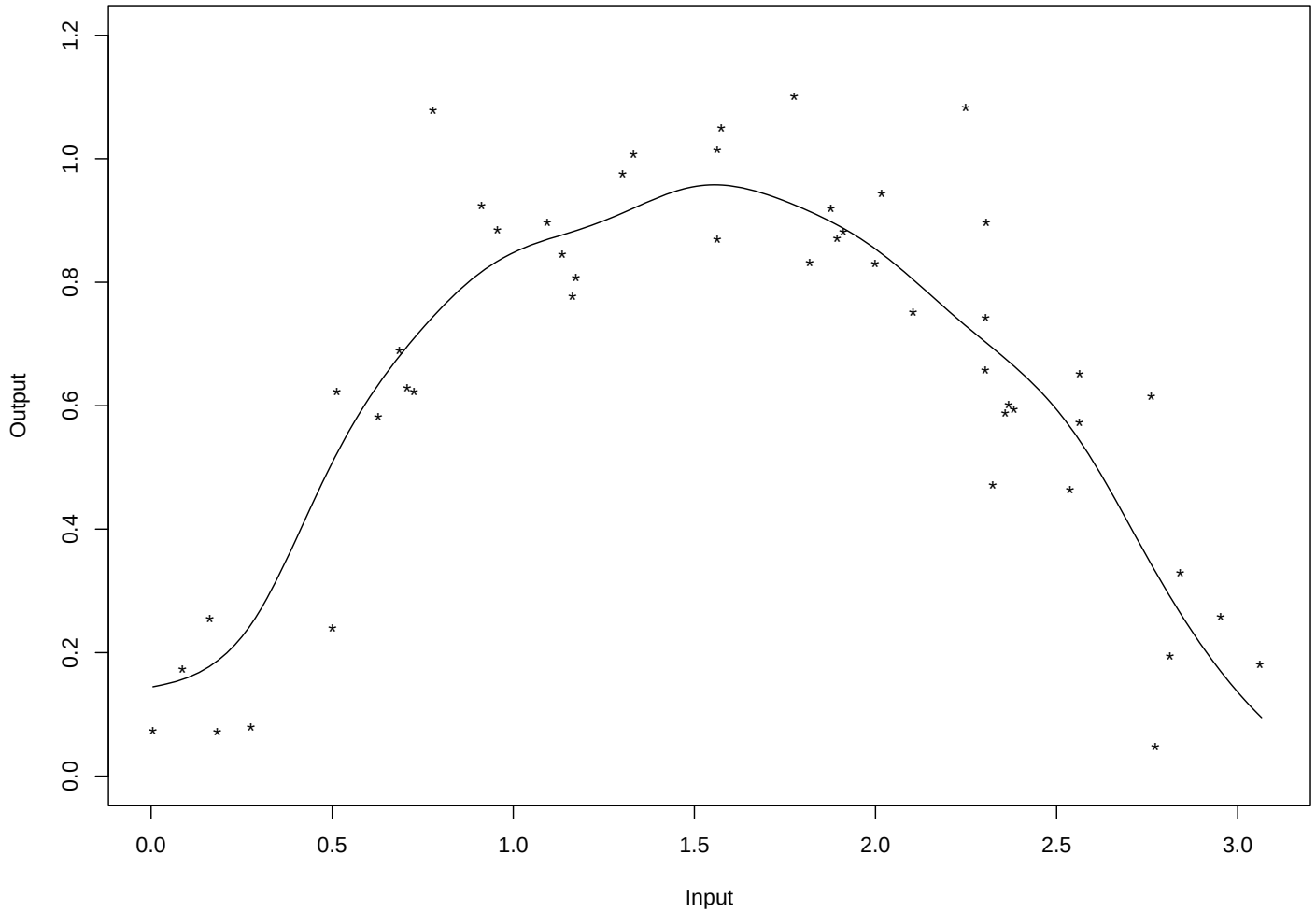
The probability of occurrence of the shingle *testimonium huic* as a function of time



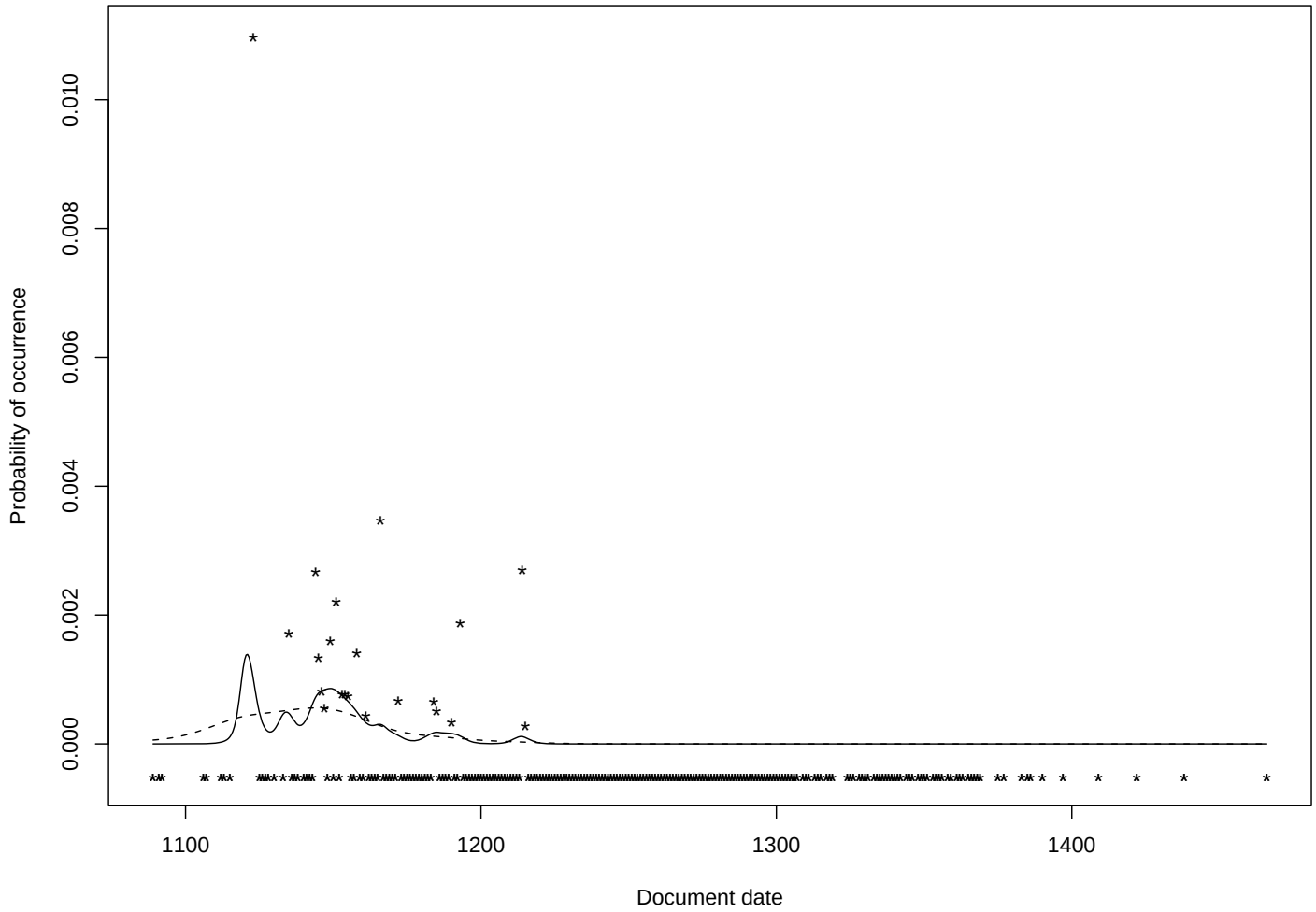

Output
Input
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Output
Input
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0


Output
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0
Input


Output
Input
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.0
0.5
1.0
1.5
2.0
2.5
3.0

The probability of occurrence of the shingle *Francis et Anglicis* as a function of time



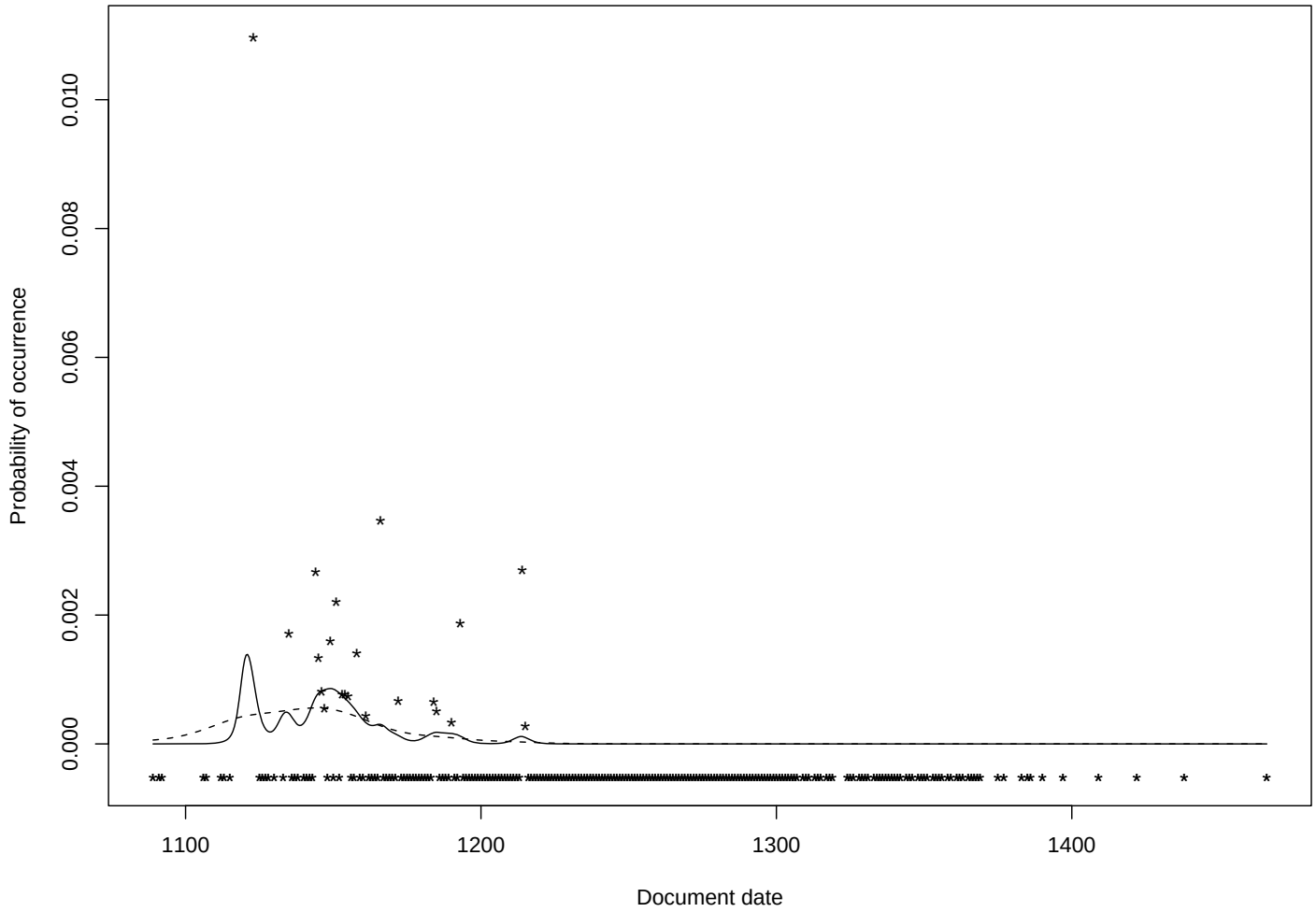
Estimating the probability of occurrences of shingles in order to date undated document $\mathcal{D}$

- Construct the set $S(\mathcal{D})$ for a fixed shingle order.

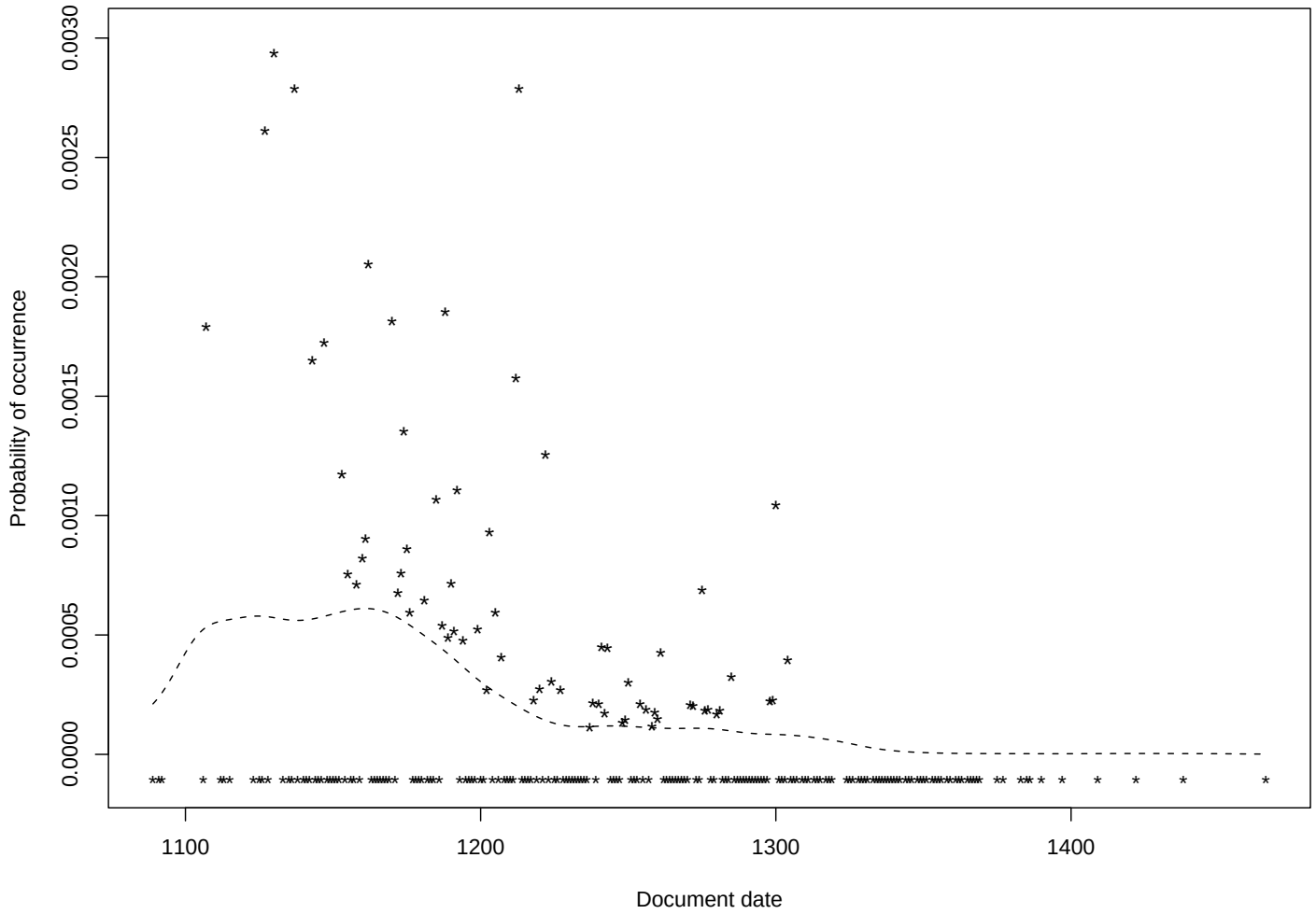
Let s_1 be *Francis et Anglicis*
Let s_2 be *ibidem Deo seruientibus*

- $P_{s_1}(1130) \times P_{s_2}(1130) \times P_{s_3}(1130) \times P_{s_4}(1130) \times \cdots$
$= 0.0007 \times 0.0005 \times \cdots$

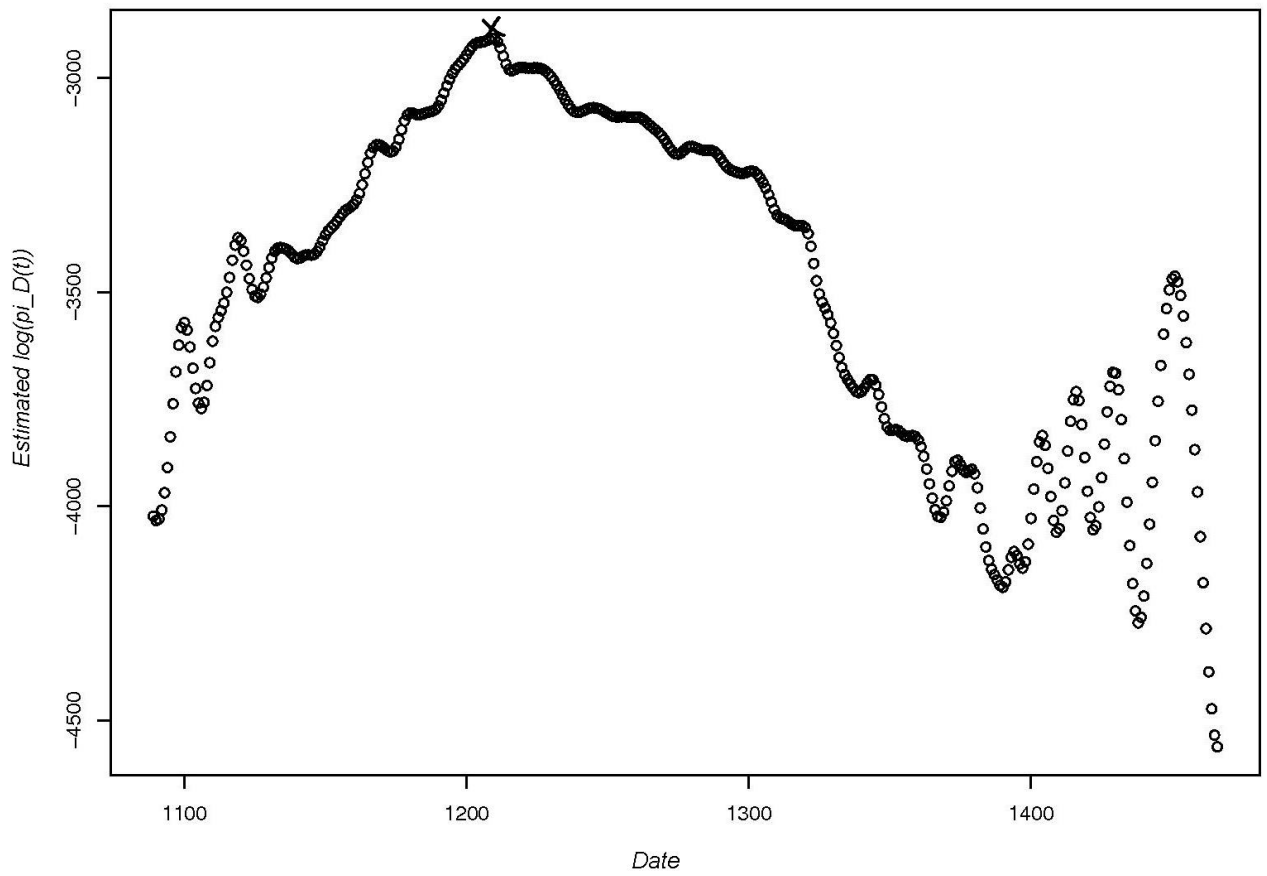
The probability of occurrence of the shingle *Francis et Anglicis* as a function of time



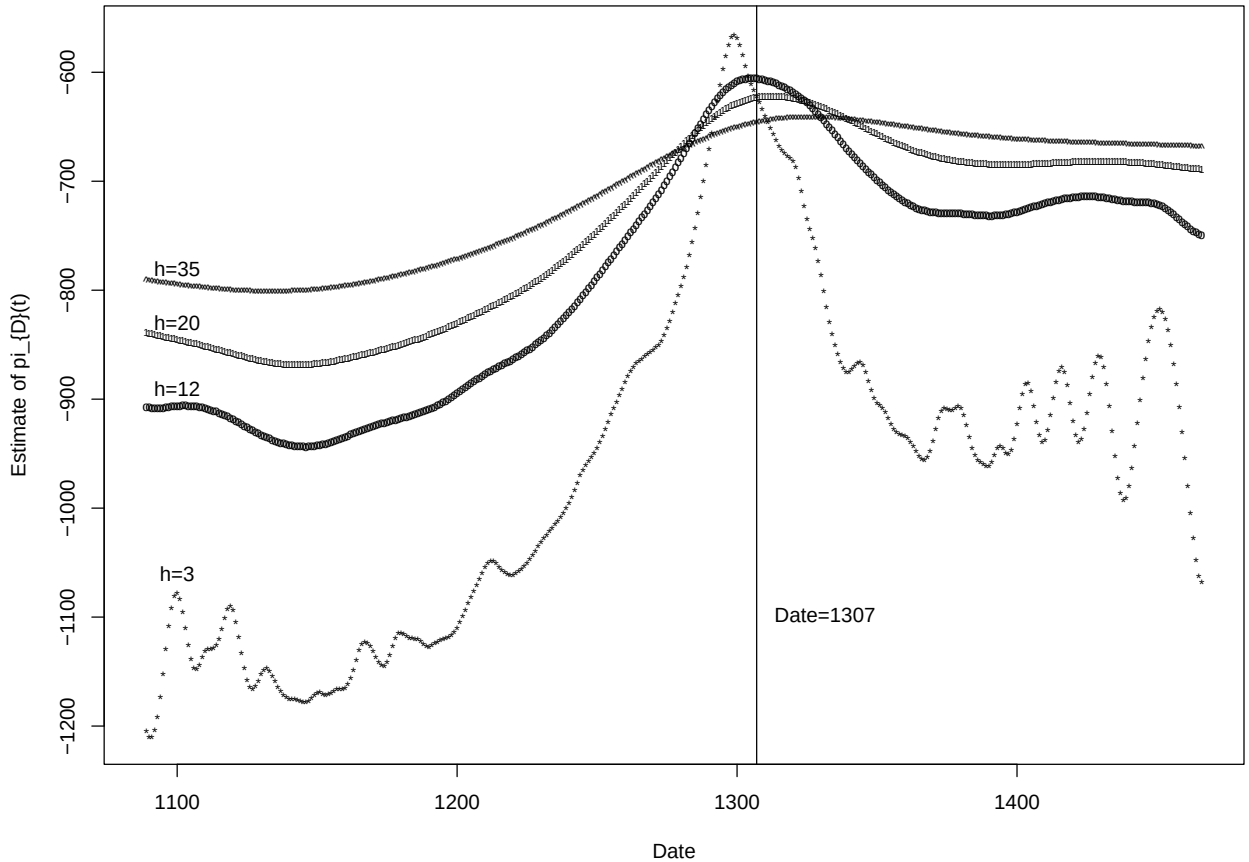
The probability of occurrence of the shingle *ibidem Deo seruientibus* as a function of time



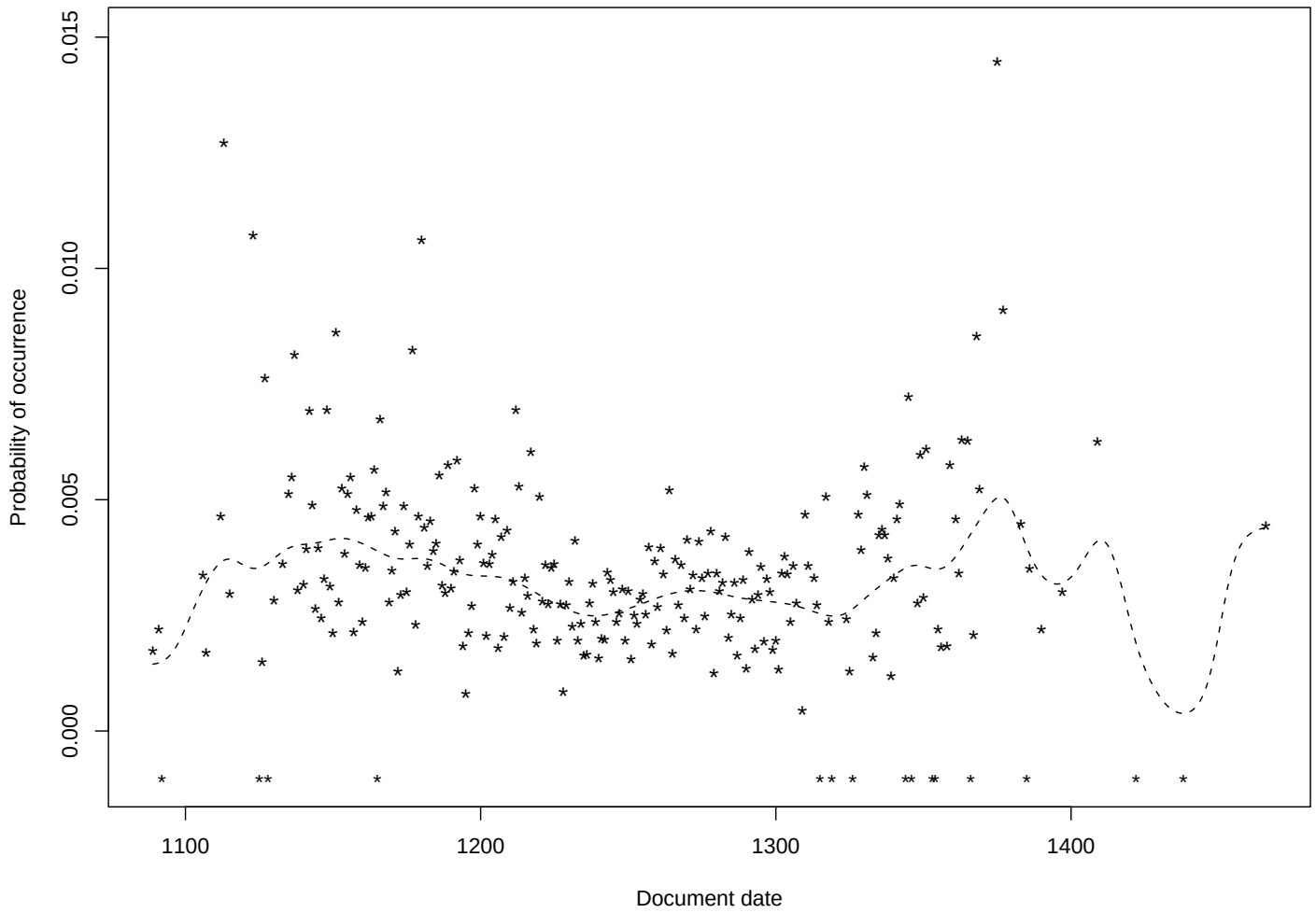
The probability of occurrence of document *D* as a function of time. The actual date for the document is 1211 and the estimated date is 1210 (the peak)



The probability of occurrence of a document as a function of time. True date is 1299. Estimated date is 1307



The Probability of occurrence of the word *omnibus* as a function of time

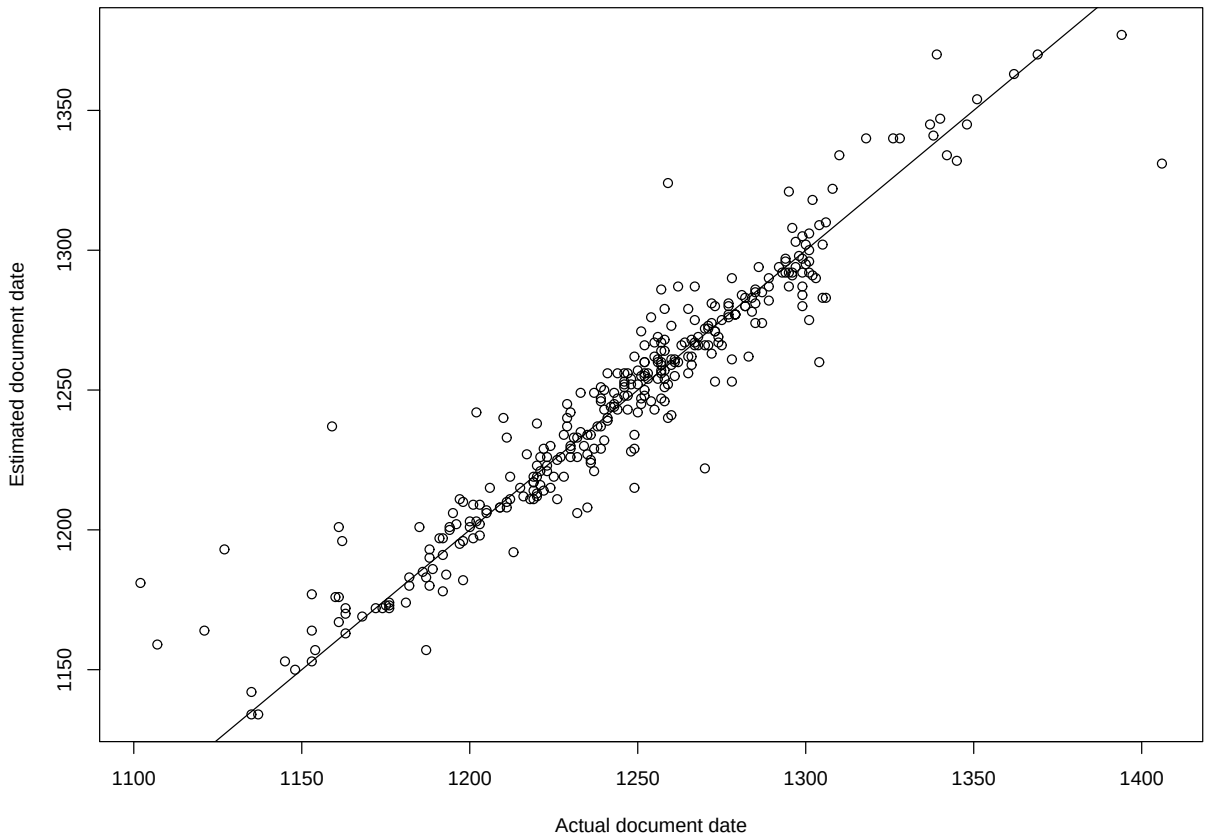


Result

Of all shingle orders, shingle order 2 performed best. On a test set of 326 documents

- Average error in absolute terms (mean) is 9.0 years
- The 50th percentile of the error in absolute terms (median) is 6.0 years

Estimated versus actual document date for the 326 documents in the test set based on shingle order 2. The solid line is “X = Y” axis.



This is joint work with Andrey Feuerverger and Gelila Tilahun

University of Toronto

Thank You!

The End