スペイン語古文書の年代・地点推定のための最適な手法の探求

川崎　義史[1] 永田　亮[2]
[1]東京大学 [2]甲南大学
ykawasaki@g.ecc.u-tokyo.ac.jp　nagata-nlp2022@ml.hyogo-u.ac.jp.

概要

現存する文献史料に立脚する歴史学や通時言語学において，文書の正確な年代推定と地点推定は最重要課題の一つである。本稿では，スペイン語古文書の作成年代と作成地点を，文書内容に基づき回帰により推定する方法を検討した。具体的には，複数の回帰モデルと文書ベクトルとの組み合わせによる比較実験を行い，それぞれの特性を分析した。実験の結果，次の 2 つの知見を得た。まず，文書ベクトルとしては，分散表現ベースのものよりも，単語・文字 *n*-gram の方が優れていた。第二に，回帰モデルとしては，加重平均 *k*-NN の性能が最良だった。また，性能はやや劣るものの，予測の信頼度として推定分散が求まるガウス過程の有効性も確認された。

1. はじめに

現存する文献史料に立脚する歴史学や通時言語学において，文書の作成年代と作成地点の正確な推定は，真贋判定とともに最重要課題の一つである。文書の作成年代と作成地点の推定は，それぞれ**年代推定**，**地点推定**と呼ばれる。作成地点を緯度・経度で表す場合，各々の推定を**緯度推定・経度推定**と呼ぶ。

文書内容に基づく年代推定と地点推定は，分類問題として解かれることが一般的である（2 節参照）。これに対し，本稿では，年代と地点を連続変数とみなし，両者の推定を**回帰問題**として扱う方法を検討した。具体的には，複数の回帰モデルと文書ベクトルとの組み合わせによる比較実験を行い，各々の特性を分析した。データとして，**中近世スペイン語古文書**を用いた。古文書とは，法令，権利，財産譲渡等に関する近代以前の手書きの行政・法律関係文書の総称である。

実験の結果，次の 2 つの知見を得た。まず，文書ベクトルとしては，分散表現ベースのものよりも，単語・文字 *n*-gram の方が優れていた。このことは，文書の全体的な情報よりも，特定の単語や文字連続が推定に効果的であることを示唆している。第二に，回帰モデルとしては，加重平均 *k*-NN の性能が最良だった。これは，内容が類似した文書が多く含まれるコーパスの性質によると考えられる。また，性能はやや劣るものの，予測の信頼度として推定分散が求まる**ガウス過程（Gaussian Process：GP）**[1]の有効性も確認された。

本稿の構成は，以下の通りである。2 節で関連研究を概観する。3 節では，データセットの説明を行う。4 節で実験設定を説明し，5 節で実験結果の報告と考察を行う。6 節でまとめと今後の課題を述べる。

2. 関連研究

文書内容に基づく年代推定と地点推定は，分類の枠組みで扱われるのが一般的である。年代推定では，一定の期間ごとに分割されたビンがクラスとなる[2]–[6]。地点推定では，州や県などの行政単位や機械的に分割された格子状の範囲がクラスとなる[6]–[9]。分類器には，ナイーブベイズ，ロジスティック回帰，サポートベクターマシン，ニューラルネットなどが用いられる。

年代推定や地点推定を回帰問題として扱う研究も僅かだが存在する。[10]は，品詞頻度を特徴量とし，芥川龍之介の作品の執筆年代を線形回帰，Support Vector Regression（SVR），ランダムフォレスト回帰を用いて推定した。[2]は，GP を用いて北欧語古文書の年代推定を行った。文書は文字・単語 *n*-gram の二値ベクトルで表現されている。[11]は，ツイートの緯度・経度をニューラル回帰で推定した。

3. データセット

データセットには，中近世スペイン語古文書コーパス CODEA+ 2015（Corpus de Documentos Españoles Anteriores a 1800）を使用した[12]。このコーパスは，1100 年から 1800 年の間にスペイン各地で発行された約 2500 文書から成る。このうち，文書長が 10 語以上で，作成年代と地点が明記されている 2076 文書の校訂版を利用した。校訂版では，省略記号が展開され，綴りや単語分割等が統一的に処理されている。

現段階では，文書の年代分布・地点分布に大きな偏りが存在する。文書数が少なく粒度の細かい推定は困難なため，作成地点の緯度・経度には，今日その地点が属する県の県都のそれを採用した。文書長の平均は 582 語, 最小値は 10 語, 最大値は 6271 語, 標準偏差は 571 語と変動が大きい。

4. 実験設定

4.1. 回帰モデル

回帰モデルとしては，GP，加重平均 *k*-NN，Ridge 回帰，SVR[13]を用いた。GP は，通常の線形回帰が適用しにくい問題に対し，より柔軟なモデリングを可能にする回帰モデルである[1]。また，予測の信頼度を分散で表現できるベイズ的なモデルでもある。カーネルには，年代推定・地点推定ともに RBF カーネルを使用した。カーネルのハイパーパラメータの最適化は，GPy で行なった[i]。最適化の容易さも GP の利点である。加重平均 *k*-NN の重みは，文書間のコサイン類似度とした。最適な$k \in [1, 10]$は，訓練データでの leave-one-out で決定した。Ridge 回帰を使用したのは, 後述のように入力次元数が大きいため, 過剰適合を緩和するためである。ハイパーパラメータはデフォルトのものを使用した[14]。SVR のカーネルには，年代推定・地点推定ともに RBF カーネルを用いた。その他のハイパーパラメータはデフォルトのものを使用した[14]。

4.2. 文書ベクトル

回帰モデルの入力となる文書ベクトルには, 単語・文字 *n*-gram の他に，Doc2Vec[15]と BERT[16]で求めた文書の分散表現を用いた。前処理として，作成年代・地点を表している箇所は人手で削除した。

単語・文字 *n*-gram は，グラムの有無の二値ベクトルと頻度ベクトルの 2 種類を使用した。$n \in \{1, 2, 3\}$とした。特徴空間は，訓練データの文書に含まれる最頻の 768，2000 パターンとした。ただし，文字 1-gram など，種類数がこれら以下になる場合がある。次元数 768 は, 後述の BERT に合わせたものである。

Doc2Vec による文書の分散表現は，gensim [17]で求めた。テストデータの文書の分散表現は，訓練データによる学習モデルから推測した。次元数は, 100, 200，768，1000 とした。その他のハイパーパラメータはデフォルトのものを使用した。

BERT による文書の分散表現には, [CLS]トークンのそれを使用した。現代スペイン語のモデル（bert-base-spanish-wwm-cased）を使用し，次元数は 768 とした[ii]。大文字・小文字の区別は無視した。入力文書長の上限（512 語）があるため，上限以上の文書は 512 語ごとに分割し，複数のサブ文書の分散表現の平均を文書の分散表現とした。

文書長が影響しないように，いずれの文書ベクトルも，ノルムが 1 になるように正規化した。正規化をしない場合は，推定性能が大幅に悪化した。

4.3. 評価指標

実験は，2076 文書を 9 対 1 で訓練データとテストデータに分割して行なった。文書数が少なく, 年代・地点分布に偏りがあるため，均質なデータ分割は難しいと判断し，開発データは準備しなかった。開発データによるハイパーパラメータの調整は，今後の課題としたい。

推定性能の評価指標には，Root Mean Squared Error（RMSE）を用いた。参考のため, Mean Absolute Error（MAE），MedAE（Median Absolute Error）も報告する。地点推定については，緯度・経度の予測値から正解地点との直線距離を GeoPy で求め，これを誤差とした[iii]。年代推定の単位は年とした。ベースラインは，訓練データの年代，緯度，経度の平均による推定とした。

5. 実験結果と考察

表 1 に各モデルと入力の組み合わせによる年代推定の実験結果を，表 2 に地点推定の実験結果を報告する。緯度推定・経度推定の結果は，付録の表 3 を参照。簡単のため，緯度・経度は同一のモデルで推定した。各モデルで最良の結果は太字にし，全体で最良のものには下線を付した。

入力が *n*-gram の場合は, 最良の結果となった *n* の場合を報告している。W-*n* と Ch-*n* は，それぞれ，単語と文字の *n*-gram を表す。末尾の-bin と-freq は，それぞれ，二値ベクトルと頻度ベクトルを表す。

図 1 は，文字 3-gram の頻度ベクトルを入力とした GP による年代（左），緯度（中央），経度（右）

[i] https://sheffieldml.github.io/GPy/
[ii] https://huggingface.co/dccuchile/bert-base-spanish-wwm-cased
[iii] https://geopy.readthedocs.io/

の予測値と正解値のプロットである。グレーの縦棒は推定標準偏差を表す。実線の対角線は，予測値が正解値と一致する場合に相当する。上下の破線は，年代推定では±50 年の区間を，地点推定では±1 度の区間を示す。地点推定では，年代が古い（新しい）文書ほど寒色（暖色）で示した。

5.1. 年代推定

表 1　年代推定の結果

モデル	入力	次元	年代誤差（年）		
			MAE	MedAE	RMSE
Baseline	*	*	161.8	164.7	183.8
GP	W-1-bin	2000	36.5	30.7	45.3
GP	Ch-3-freq	2000	**29.5**	**22.7**	**38.5**
GP	Doc2Vec	768	61.3	53.4	74.8
GP	BERT	768	37.0	30.6	47.4
k-NN	W-1-bin	2000	**24.2**	**17.5**	36.3
k-NN	Ch-3-bin	2000	24.6	19.1	**33.7**
k-NN	Doc2Vec	100	31.1	20.7	44.3
k-NN	BERT	768	41.9	33.2	54.5
Ridge	W-1-bin	2000	37.9	29.6	48.7
Ridge	Ch-3-bin	2000	**32.5**	**23.1**	**41.6**
Ridge	Doc2Vec	768	58.0	51.1	72.7
Ridge	BERT	768	51.2	44.8	62.5
SVR	W-1-bin	768	138.1	140.7	165.2
SVR	Ch-3-freq	768	120.6	120.5	147.3
SVR	Doc2Vec	768	**99.4**	**91.7**	**119.7**
SVR	BERT	768	130.2	133.9	158.0

推定性能は *k*-NN が最良となり，続いて GP，Ridge 回帰，SVR の順となった。いずれのモデルもベースラインを上回った。*k*-NN の性能が良いのは，内容が類似した文書が多く含まれるコーパスの性質によると考えられる。非線形の関係も捉えられる GP は，より単純なモデルの *k*-NN を上回らなかった。今後，複数のカーネルの組み合わせ等を検討する必要がある。Ridge 回帰の性能も悪くはなく，線形に変化する変数の存在が示唆される。SVR は，他のモデルよりも大幅に低い結果となった。

入力は，Doc2Vec や BERT による分散表現ベースのものよりも，*n*-gram の方が優れていた。これは，文書の全体的な情報よりも，*n*-gram が捉える具体的な単語や文字連続が推定に効果的であることを示唆している。全般的に，単語よりも文字 *n*-gram の方が優れていた。どのモデルでも単語では$n = 1$，文字では$n = 3$での性能が最良だった。SVR 以外は，次元数が大きい方が高性能だった。二値ベクトルと頻度ベクトルの優劣は，モデルに依存するようである。

Doc2Vec と BERT の分散表現の優劣は判然としなかった。使用した BERT モデルは現代スペイン語のものだが，中近世語に対しても，ある程度，有効なようである。これは，BERT が subword を考慮するため未知語にも対応可能であり，また，現代語と中近世語の乖離が小さいためだと考えられる。

GP の性能はやや劣るが，予測の信頼度として推定標準偏差が求まることは大きな利点である。モデルと入力の種類によらず，推定標準偏差と絶対値誤差には 0.2 程度の正の相関が見られた（付録の図 2）。

モデルと入力の種類によらず，文書長が大きいほど，誤差が小さくなる傾向が見られた（図は非掲載）。長い文書ほど，推定に寄与する語が多く含まれるためだと考えられる。地点推定でも同様の傾向が見られた。

5.2. 地点推定

年代推定と同様に，*k*-NN が最良のモデルとなり，続いて GP，Ridge 回帰，SVR の順となった。いずれのモデルもベースラインを上回った。年代推定と異なり，SVR も他のモデルと匹敵する性能だった。入力は，年代推定同様，*n*-gram の性能が良かった。GP では，推定標準偏差と絶対値誤差には相関は見られなかった（付録の図 2）。

緯度・経度の RMSE を各々のレンジ（最大値−最小値）で割り標準化すると，経度の方が緯度よりも誤差が 20%程度小さくなった。つまり，経度の方が緯度よりも推定しやすいことを意味する。これは，中世スペインの再征服運動の結果，東西方向よりも南北方向の言語的特徴が類似していることの反映だと考えられる[18]。

年代が新しい文書は，経度推定の誤差が大きくなった（図 1）。これは，年代の新しい文書の少なさに加え，16 世紀以降，半島西部・中部・東部の方言的特徴が消失し言語が均質化したため，経度の推定が困難になったためだと考えられる[19]。

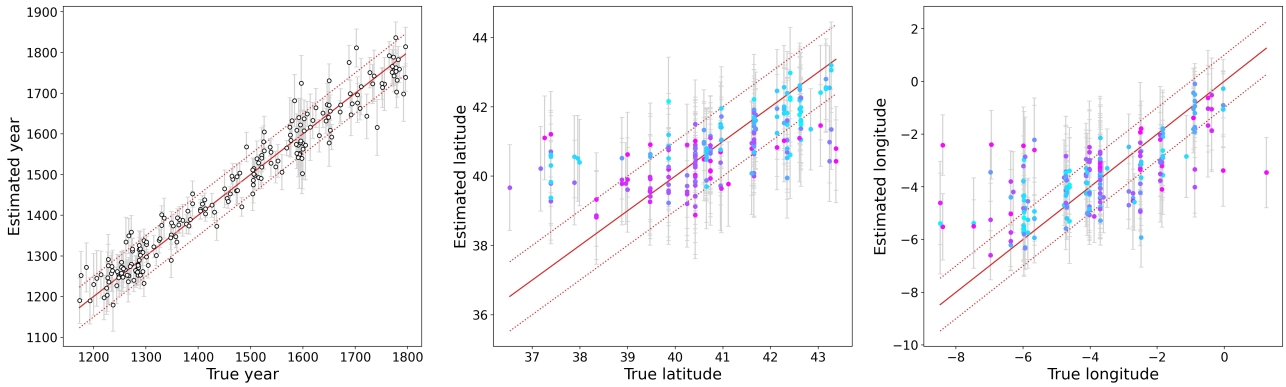

図 1　文字 3-gram の頻度ベクトルを入力とした GP による年代（左），緯度（中央），経度（右）の予測値と正解値のプロット。グレーの縦棒は推定標準偏差を表す。実線の対角線は，予測値が正解値と一致する場合に相当する。上下の破線は，年代推定では±50 年の区間を，地点推定では±1 度の区間を示す。地点推定では，年代が古い（新しい）文書ほど寒色（暖色）で示した。

表 2　地点推定の結果

モデル	入力	次元	地点誤差（km）		
			MAE	MedAE	RMSE
Baseline	*	*	201.2	170.6	233.9
GP	W-2-bin	2000	140.5	102.3	178.6
GP	Ch-3-freq	2000	**137.4**	**102.0**	**174.2**
GP	Doc2Vec	1000	159.4	126.7	197.4
GP	BERT	768	155.8	127.5	192.6
k-NN	W-1-bin	2000	**118.4**	**80.3**	169.7
k-NN	Ch-3-bin	2000	124.5	93.6	**169.4**
k-NN	Doc2Vec	200	143.5	112.6	186.6
k-NN	BERT	768	156.4	126.6	196.1
Ridge	W-1-bin	2000	**138.8**	**100.3**	181.4
Ridge	Ch-3-freq	2000	143.6	116.0	**179.0**
Ridge	Doc2Vec	100	163.9	125.0	204.1
Ridge	BERT	768	170.1	144.0	204.8
SVR	W-1-bin	2000	**134.8**	**83.6**	184.6
SVR	Ch-3-freq	2000	136.7	99.2	**181.1**
SVR	Doc2Vec	1000	151.5	113.4	194.4
SVR	BERT	768	165.1	131.1	206.5

6. おわりに

本稿では，スペイン語古文書の作成年代と作成地点を，文書内容に基づき回帰により推定する方法を検討した。具体的には，複数の回帰モデルと文書ベクトルとの組み合わせによる比較実験を行い，それぞれの特性を分析した。実験の結果，次の 2 つの知見を得た。まず，文書ベクトルとしては，分散表現ベースのものよりも，単語・文字 *n*-gram の方が優れていた。第二に，回帰モデルとしては，加重平均 *k*-NN の性能が最良だった。また，性能はやや劣るものの，予測の信頼度として推定分散が求まるガウス過程の有効性も確認された。

今後の課題は，以下の 4 点である。まず，カーネルや入力ベクトルの次元数などのハイパーパラメータの調整が必要である。この点で，GP はハイパーパラメータが自動的に求まるという利点がある。*n*-gram を入力とする場合，中頻度層のものも考慮することで，性能向上が期待される。第二に，年代推定と地点推定をマルチタスクとして扱い，ニューラルに回帰することが考えられる。緯度と経度には相関がないが，年代と緯度，年代と経度には相関がある。これらの相関を考慮し，年代と地点を同時に予測することで性能向上が期待される[6]。第三に，同一の入力を使用した分類モデルとの比較が必要である。第四に，[6], [20]のように，推定の根拠となる単語・文字列の特定が望まれる。*n*-gram を入力とする場合，重みの絶対値の大きなものの特定は容易である。GP では，関連度自動決定により，推定に影響のある次元の特定ができる[1]。一方，Doc2Vec や BERT による分散表現を入力とする場合，ある次元が具体的な語に対応しているわけではないので，特定は難しいように思われる。

謝辞

本研究はJSPS科研費18K12361の助成を受けたものです。

参考文献

[1] 持橋大地，大羽成征，『ガウス過程と機械学習』. 東京: 講談社, 2019.

[2] S. Boldsen and F. Wahlberg, “Survey and Reproduction of Computational Approaches to Dating of Historical Texts,” in *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, 2021, pp. 145–156.

[3] F. De Jong, H. Rode, and D. Hiemstra, “Temporal Language Models for the Disclosure of Historical Text,” in *Humanities, Computers and Cultural Heritage: Proceedings of the 16th International Conference of the Association for History and Computing (AHC 2005)*, 2005, pp. 161–168.

[4] G. Tilahun, A. Feuerverger, and M. Gervers, “Dating Medieval English Charters,” *Ann. Appl. Stat.*, vol. 6, no. 4, pp. 1615–1640, 2012.

[5] N. Kanhabua and K. Nørvåg, “Using Temporal Language Models for Document Dating,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 5782 LNAI, no. PART 2, pp. 738–741, 2009.

[6] Y. Kawasaki, “Dating and Geolocation of Medieval and Modern Spanish Notarial Documents Using Distributed Representation,” in *QUALICO2021 Book of Abstracts*, 2021, pp. 34–36.

[7] S. Roller, M. Speriosu, S. Rallapalli, B. Wing, and J. Baldridge, “Supervised Text-Based Geolocation Using Language Models on an Adaptive Grid,” in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 1500–1510.

[8] B. Han, P. Cook, and T. Baldwin, “Text-Based Twitter User Geolocation Prediction,” *J. Artif. Intell. Res.*, vol. 49, pp. 451–500, 2014.

[9] B. Wing and J. Baldridge, “Hierarchical Discriminative Classification for Text-Based Geolocation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 2014, pp. 336–348.

[10] 金明哲, “文章の執筆時期の推定：芥川龍之介の作品を例として,” 行動計量学, vol. 36, no. 2, pp. 89–103, 2009.

[11] P. Mishra, “Geolocation of Tweets with a BiLSTM Regression Model,” in *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 2020, pp. 283–289.

[12] GITHE (Grupo de Investigación Textos para la Historia del Español), “CODEA+ 2015 (Corpus de Documentos Españoles Anteriores a 1800).” [Online]. Available: http://www.corpuscodea.es/.

[13] C. C. Chang and C. J. Lin, “LIBSVM: A Library for Support Vector Machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[14] F. Pedregosa *et al.*, “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[15] Q. Le and T. Mikolov, “Distributed Representations of Sentences and Documents,” in *Proceedings of the 31st International Conference on Machine Learning*, 2014, vol. 32, no. 2, pp. 1188–1196.

[16] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019, pp. 4171–4186.

[17] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.

[18] R. Penny, *A History of the Spanish Language*. Cambridge: Cambridge University Press, 2002.

[19] R. Penny, *Variation and Change in Spanish*. Cambridge: Cambridge University Press, 2000.

[20] A. Rahimi, T. Baldwin, and T. Cohn, “Continuous Representation of Location for Geolocation and Lexical Dialectology using Mixture Density Networks,” in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 2017, pp. 167–176.

付録

表 3　緯度推定・経度推定の結果

モデル	入力	次元	緯度誤差（度）			経度誤差（度）			地点誤差（km）		
			MAE	MedAE	RMSE	MAE	MedAE	RMSE	MAE	MedAE	RMSE
Baseline	*	*	1.20	1.05	1.52	1.51	1.13	1.92	201.2	170.6	233.9
GP	W-2-bin	2000	0.85	**0.55**	1.19	1.03	0.72	1.43	140.5	102.3	178.6
GP	Ch-3-freq	2000	**0.82**	0.57	**1.15**	**1.02**	**0.71**	**1.40**	**137.4**	**102.0**	**174.2**
GP	Doc2Vec	1000	0.99	0.73	1.32	1.14	0.81	1.57	159.4	126.7	197.4
GP	BERT	768	0.94	0.74	1.29	1.14	0.88	1.53	155.8	127.5	192.6
k-NN	W-1-bin	2000	**0.73**	**0.37**	1.15	**0.84**	**0.52**	**1.33**	**118.4**	**80.3**	169.7
k-NN	Ch-3-bin	2000	0.75	0.47	**1.14**	0.90	0.60	1.33	124.5	93.6	**169.4**
k-NN	Doc2Vec	200	0.85	0.55	1.21	1.07	0.79	1.53	143.5	112.6	186.6
k-NN	BERT	768	0.94	0.62	1.32	1.14	0.81	1.55	156.4	126.6	196.1
Ridge	W-1-bin	2000	**0.81**	**0.48**	1.19	**1.06**	**0.74**	1.47	**138.8**	**100.3**	181.4
Ridge	Ch-3-freq	2000	0.83	0.59	**1.17**	1.11	0.88	**1.46**	143.6	116.0	**179.0**
Ridge	Doc2Vec	100	0.98	0.70	1.33	1.23	0.89	1.67	163.9	125.0	204.1
Ridge	BERT	768	0.96	0.68	1.32	1.34	1.15	1.70	170.1	144.0	204.8
SVR	W-1-bin	2000	0.79	0.43	1.24	**1.02**	**0.67**	**1.45**	**134.8**	**83.6**	184.6
SVR	Ch-3-freq	2000	**0.77**	**0.42**	**1.18**	1.07	0.77	1.48	136.7	99.2	**181.1**
SVR	Doc2Vec	1000	0.93	0.61	1.30	1.12	0.76	1.55	151.5	113.4	194.4
SVR	BERT	768	0.91	0.58	1.34	1.31	1.14	1.70	165.1	131.1	206.5

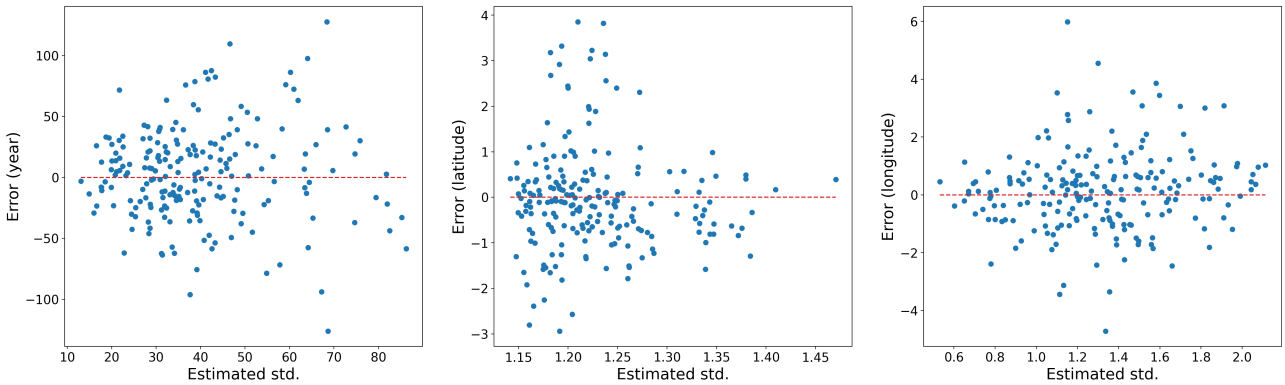

図 2　文字 3-gram の頻度ベクトルを入力とした GP による推定標準偏差と年代推定（左），緯度推定（中央），経度推定（右）の誤差のプロット。年代推定では，推定標準偏差と絶対値誤差には 0.2 程度の正の相関が見られた。地点推定では，相関は見られなかった。