Short Term E-commerce Sales Forecast Method Based on Machine Learning Models

Tingli Feng[1], Chenming Niu[2], and Yuchen Song[2(✉)]

[1] College of Engineering and Applied Science, University of Wisconsin-Milwaukee, Milwaukee 53211, US
[2] School of Transportation, Southeast University, Nanjing 21008, China
213193120@seu.edu.cn

Abstract. Nowadays, e-commerce is developing rapidly in the world. In 2010, China's e-commerce turnover reached 37.21 trillion yuan. For modern e-commerce corporations, an accurate sales forecast is the key to driving the development of corporations. While many effective forecast methods have been established in multiple business contexts, few of these methods have achieved good results in the short-term forecast and the value of detailed classified information of promotional plans has not yet been explored. This study attempts to establish a short-term forecast framework and explore whether incorporating detailed promotional plans can improve the forecast accuracy of the forecasting framework established. This study proposes a short-term forecast framework and implements six machine learning models to forecast daily sales. It finds that in a short-term forecast with one month's data, the framework proposed can achieve rather good performance with out-of-sample MAPE ranging from 10.23% to 20.83% in different machine learning models. The incorporation of the detailed classification of discount information results in statistically significant improvements in the out-of-sample accuracy of linear regression, ridge regression, and lasso regression, with the best improvement of 36.19% in MAPE, but has no significant influence on the support vector machine, gradient boosting and random forest. From these results, the study provides recommendations for short-term forecast sales in general as well as a detailed classification of discount information.

Keywords: Short-term forecast · Machine learning · Discount information

1 Introduction

1.1 Background

Today, e-commerce is in more and more increasing competition, so companies must adopt strategic planning and make the right marketing decision. The first step in the planning and decision-making process is to predict the future demand for products and thus the resources can be adjusted in time to meet the demand [1]. The importance of

Tingli Feng, Chenming Niu and Yuchen Song—These authors contributed equally

G. Ali et al. (Eds.): ISEMSS 2022, ASSEHR 687, pp. 1020–1030, 2022.
https://doi.org/10.2991/978-2-494069-31-2_119

forecasting to corporations has been discussed by many authors and experts in the field [2, 3]. Taking Walmart as an example, using machine learning models for sales forecast from massive historical data helps optimize their business operation, ranging from cash flow, and stall to production and financial management [4]. Moreover, it can reduce uncertainty and anticipate change in the market.

Many machine learning (ML) and deep learning models, such as Support Vector Machine (SVM), Random Forest (RF), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost) [5, 6], have achieved rather good performance in the sales forecast. However, these machine learning models require inputting a large volume of data to find the connections in the data and get a well-performed function. As a result, these models are often trained with months or years of data and few researchers can get a good performance using days of data to forecast sales.

As for sales information, many companies forecast sales with a single sale attribute, a simple promotional plan, and seasonal factors [7]. Few companies refine the discount information and its value of it is unknown [8]. Most customers can get different kinds of discounts through bundle reduction, direct reduction, coupon reduction, and so on. And their purchasing decisions are greatly influenced by different kinds of discounts. Several academic studies in marketing have already demonstrated the value of discount information on strategy making and attracting consumers. Despite the attention and research that have been devoted to discount information, the value of the different kinds of discounts on sales forecasts has not yet been well studied.

1.2 Related Work

Hassan et al. pointed out a reinforcement learning framework that improves the efficiency and accuracy of demand forecasting on a rolling horizon. The method's benefit is that it can be used for most time series and machine learning models. The demand forecasting technique relies on a rolling horizon, which is updated (rolled) at the end of each week when new information becomes available. Then, the model is updated with the most recent information to improve the accuracy [9].

Cui et al. developed seven linear and nonlinear machine models to find out the effect of social media in increasing the accuracy of daily sales forecasts. After comparing models with variable selection and without variable selection to calculate the out-of-sample MAPE as an evaluation, The best-performing method is the random forest, and adding social media information can improve the forecasts [10].

Hasselbeck et al. emphasized that forecasting short-term and long-term demand is the ultimate necessity for operations management. They predicted the Horticultural Sales with more than eleven methods with classical forecasting and machine learning including Lasso Regression, Ridge Regression, Long Short-Term Memory Network (LSTM), Seasonal Autoregressive Integrated Moving Average (SARIMA), XGBoost, and so on. As a result, XGBoost is the top performer which has the lowest Mean Absolute Percentage Error (MAPE) [11].

Falatouri et al. emphasized the importance of predictive analytics (PA) in demand forecasting of Supply Chain Management (SCM). They compared the SARIMA with LSTM models in stable and seasonal demand and the external factor of promotions.

The two models have good forecasting performances. Furthermore, they combined two models to adapt to more situations [12].

Ensafi et al. compared many machine learning models to do the forecasting with time-series forecasting of seasonal item sales such as SARIMA and Triple Exponential Smoothing, Prophet, (LSTM), and Convolutional Neural Network (CNN) to find better performances of models. As a result, the Prophet and CNN have the lower Root Mean Squared Error (RMSE) and MAPE [13].

1.3 Research Methodology

This study aims to establish a short-term forecast framework with ten days' data as the training set and the next day as the testing set and apply six machine learning models in the framework to evaluate the accuracy and robustness of this framework and examine the influence and value of discount information on sale forecast.

The Data set in this study incorporates the detailed discount information which is mainly divided into three categories: 'direct discount', 'quantity discount' and 'bundle discount', and a one-month sales record for a product of one category. The Dataset is obtained through the '2020 MSOM Data-Driven Research Challenge'. The dataset describes 2.5 million customers (457,298 made purchases) and 30,000 SKUs (from one product category) during March 2018.

This study proposes a short-term forecast framework with ten days' data as the training set and the next day's data as the testing data. And implement six machine learning models with different features. The models used include simple linear regression, ridge regression, lasso regression, SVM with the poly kernel, GB, and RF, among which lasso regression, GB, and RF are the models with parameter selection, and the rest three models without selection. These ML models are applied in two separate aggregate daily sales forecasts: (1) a base forecast that uses only sales information. (2) an advanced forecast that incorporates discount information which is expressed as a discount level (level 1 to 5 with a higher level meaning more price reduction) in our study. In each type of forecast, cross-validation is implemented to select the hyperparameters of GB and RF. The comparison between out-of-sample forecast accuracy using these two types of forecasts can quantify the value of discount information in the short-term forecast framework.

2 Forecast Framework and ML Models

2.1 Data Aggregation

This study aims to forecast overall_sale on a daily scale, the transactional data cannot support our research. The transactional data is aggregated on a daily scale. The specific method of aggregate is shown in Table 1.

2.2 Forecast Framework

Two types of forecasts are constructed: "base forecast" which only includes daily overall sale features and a more complete model called "advanced forecast" which includes both

Table 1. The Specific Method of Aggregate

Original data	index	Aggregate method
Quantity	t	/
Original price	op	/
Final price	fp	$\sum f*t$
Direct discount	dd	$\sum dd*op*t / \sum op*t$
Quantity discount	qd	$\sum qd*op*t / \sum op*t$
Buddle discount	bd	$\sum bd*op*t / \sum op*t$
Overall discount	od	$\sum fp*t / \sum op*t$

daily overall sale features and daily discount features. Every machine learning model is fitted for these two types of forecasts and compares their accuracy where the only difference between them is whether discount features are included.

2.2.1 Base Forecast Framework

Discount features are not included in the base forecast. In the base forecast, overall sale on day t is the function of the overall sales in the past five days[3].

$$S_t^{base} = f_i(S_{t-1}, S_{t-2}, \ldots\ldots S_{t-5}) \tag{1}$$

S_t represents the overall sale on day t. $f_i(\cdot)$ represents the different machine learning models chosen.

2.2.2 Advanced Forecast Framework

The advanced forecast includes discount features. In the advanced forecast, overall sale on day t is the function of the overall sale, and discount features in the past five days.

$$S_t = f_i(S_{t-1}, S_{t-2}, \ldots S_{t-5}, D_{t-1}, D_{t-2}, \ldots D_{t-5}) \tag{2}$$

S_t represents the overall sale on day t. D_t represents the discount features on day t such as quantity discount, direct discount, and overall discount. $f_i(\cdot)$ represents the different machine learning models chosen.

2.3 Training, Cross-Validation, and Out-of-Sample Evaluation

In-sample training data and out-of-sample testing data are split to evaluate out-of-sample accuracy. To be specific, training data and testing data are split in a rolling mechanism.

Assuming that day_t is the testing data, its corresponding training data is:

$$\text{Train set} = [\text{day}_{t-1},\ \text{day}_{t-2},\ \text{day}_{t-3}, \ldots, \text{day}_{t-m}] \tag{3}$$

Table 2. Machine learning models

Without parameter selection	With parameter selection
Linear regression	Random tree
Ridge regression	Gradient boosting
	Lasso regression

In Eq. (3), m = 10.

Assuming that the index of rolling is i, then for rolling i, the testing data and training data are shown below:

$$Train\ set = [day_i,\ day_{i+1},\ day_{i+2},\ \ldots day_{i+9}] \tag{4}$$

$$\text{Test set } = \left[\text{day}_{\text{i+10}}\right] \tag{5}$$

Cross-validation is used to select hyperparameters in machine learning models. Ten-fold cross-validation with five repeats is used to evaluate the performance of each hyperparameter. The training set is randomly divided into ten subsets of the same size, nine sets are used to train for hyperparameter selection and 1set is used to test the performance of hyperparameters. Each subset will be treated as testing set at least once.

The overall performance of the model is the average of ten subsets. Next, retain the hyperparameters with the best performance and then estimate parameters with the entire training set. After doing this, the result is to get the best model for each training set.

Out-of-sample evaluation. The best hyperparameter and parameter selected are retained for each training set during the process of constructing the forecast for the out-of-sample testing set. When forecasting the overall sale on day t, pass data and models are used from $\text{day}_{\text{t-10}}$ to $\text{day}_{\text{t-1}}$ as input to the selected best model and then get the forecast result. To be more specific, in the next round of forecasting on $\text{day}_{\text{t+1}}$, pass data and models from $\text{day}_{\text{t-9}}$ to day_{t} are carried out. This kind of rolling update mechanism have relatively good performance in small-scale data with high dimension.

2.4 ML Models

A variety of machine learning models are adopted in this study. Each of our training sets has ten days whereas our advanced forecast needs to estimate 230 parameters (23 per day times 10 days time scale). The number of independent variables is much larger than the scale of data. In this case, several machine learning models applied contain advanced parameter selection, such as random forest and gradient boosting.

SVM with the poly kernel and linear models (linear regression, lasso regression, and ridge regression) are for their low computational consumption. Next, briefly summarize each machine learning model. The ML model adopts shown in Table 2.

Random forest uses the idea of ensemble learning to ensemble multiple decision trees, the elementary decision unit of random forest is the regression tree. The regression tree divides the feature vector X into several regions which don't overlap with each other and use the mean value of each region to label the region. The loss function of the regression tree is $Loss(y, f(x)) = (f(x) - y)^2$. In the random forest, every decision tree is a classifier, the outputs of all decision trees are used in voting, and the output with the most votes is the result of the random forest [14]. The most important advantage of random tree is high accuracy. However, the overall forecast process of random forest contains two parts: parameter selection and fit process, parameter selection section consumes a relatively longer time than the fit section, which may be up to hours.

Gradient boosting is a kind of machine learning model using gradient descent.

The basic principle of gradient boosting is to use the gradient of the current model's loss function to train the newly added machine learning model [15].

SVR (Support Vector Regression), is a kind of machine learning model with a linear fitting function $y = ax + b$. However, SVR is quite different from traditional linear regression models. SVR create an interval zone on both side of the fitting function and create two slack variable α and β to set more data dot into this interval zone as possible and use these data dot inside the interval zone to calculate the loss function. In this research, SVM of the poly kernel is applied in forecast sales [16].

Linear regression. It is the simplest linear regression to forecast with all the features in JD.com data even if some of these features are not significant [17].

Lasso regression. The difference between lasso regression and traditional linear regression is in its loss function.

$$\min n^{-1} * \sum \left(y_i - \beta_0 - x_i^T * \beta \right)^2 \tag{6}$$

$$subjectto \beta^2 < t \tag{7}$$

Lasso is especially suitable for data set to use high dimensional feature set and relatively small data set size, so in this study, lasso regression is considered [18].

Ridge regression can solve the 'Multicollinearity problem' when the least square method is considered. Ridge regression adds a Regularization factor into the original objective function to restrict the outcome of the least square method. This kind of method is similar to lasso regression, the difference between ridge regression and lasso regression is that the regularization factor of lasso regression is $\lambda * |\omega|$ and the regularization factor of ridge regression is $\lambda * |\omega|^2$ [19, 20].

2.5 Data

The Data set provides transaction-level data for March 2018 during which there were no major holidays or promotions. The data set includes anonymized key identification information such as user ID and Stock Keeping Unit (SKU) ID. Each SKU can be

Table 3. Data Specification

Field	Data type	Description
Order ID	string	Order unique identification code
User ID	string	User unique identification code
SKU ID	string	SKU unique identification code
Order date	string	Order date (format: yyyy-mm-dd)
quantity	int	Number of units ordered
Type	int	1P or 3P orders
Original price	float	Original list price
Final price	float	Final purchase price
Direct discount	float	Discount due to SKU direct discount
Quantity discount	float	Discount due to purchase quantity
Bundle discount	float	Discount due to "bundle promotion"
Coupon discount_	float	Discount due to customer coupon

identified either as "*first party owned*" (1P) or "*third-party owned*" (3P), depending on the ownership of the inventory of that SKU. The detail of our data can be seen in Table 3.

3 Result Analysis and Discussion

3.1 Result

Six machine learning models are implemented, to predict overall sales with and without discount information. First, the performance of different models with discount information in the short-term forecast framework is compared to evaluate the performance of the framework in different models. Then fifteen out-of-sample MAPE are obtained for each ML model. Figure 1 shows a box diagram of the out-of-sample MAPE for these models. In the basic forecast framework, the average MAPE and MAPE distribution variance of RF is the best with an average MAPE of 10.7% and SVM is the worst one with an average MAPE of 19.31%. All six models have a relatively small variance in the distribution of the out-of-sample MAPE, so it's reasonable for us to set the mean value as the standard to evaluate the performance of the base forecast framework with different ML models. The small variance also indicates that the forecast result is relatively stable without much fluctuation, demonstrating the robustness of the base forecast framework.

Compare the performance between pairs of forecasts with and without discount information to evaluate the value of discount information. Table 2 summarizes the mean value of out-of-sample MAPE for each of the ML models. Figure 2 also shows the mean value of out-of-sample MAPE for each model in two different forecasting frameworks. Both the table and the figure show that the discount information is valuable to forecast (Table 4).

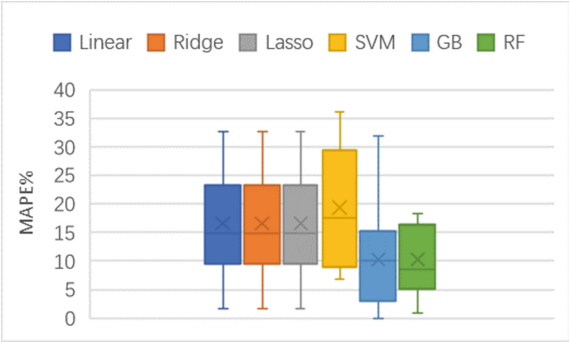

Fig. 1. The box diagram of MAPE% in the base forecast framework with six ML models

Table 4. Comparison of Out-of-Sample MAPE% for Statistical Learning Models

	Linear	Ridge	Lasso	SVM	GB	RF
Advanced Forecast Framework (with discount information)	14.91	14.92	10.55	20.83	10.71	10.70
Base Forecast Framework (without discount information)	16.54	16.54	16.54	19.31	10.31	10.22

A T-test is applied in the out-of-sample MAPE of two frameworks and the improvements of discount information with linear, ridge, and lasso are significant and the differences between SVM, GB, and RF are insignificant. From Table 2 and Fig. 2, the result indicates that the advanced forecast framework with discount information can improve the performance of the three traditional linear models and do poor in SVM and the two ensemble nonlinear models.

3.2 Discussion

Figure 2 shows that the advanced forecast incorporating discount information is superior to the base forecast using Linear, Ridge, Lasso, and GB, further confirming the finding that discount information is valuable in the forecasting framework proposed using certain ML models. The RF and GB perform well in both the base forecast and the advanced forecast, consistent with their reputation of the strong capability to deal with high dimensional data and achieve high accuracy.

SVM performs the worst in both forecast frameworks for the reason that the data used in both two forecast frameworks are high dimensional and small volume. The small volume of data indicates less information about one feature, which is hard for SVM to capture all the features' information and easy for SVM to underfit. Compared with the base forecast, the advanced forecast comes with a higher dimension but without more data, causing the lower accuracy of SVM.

In addition, Linear, Ridge, and Lasso all perform poorly in the base forecast but the accuracy of Lasso improves largely by 36.19% compared to the poor improvement of

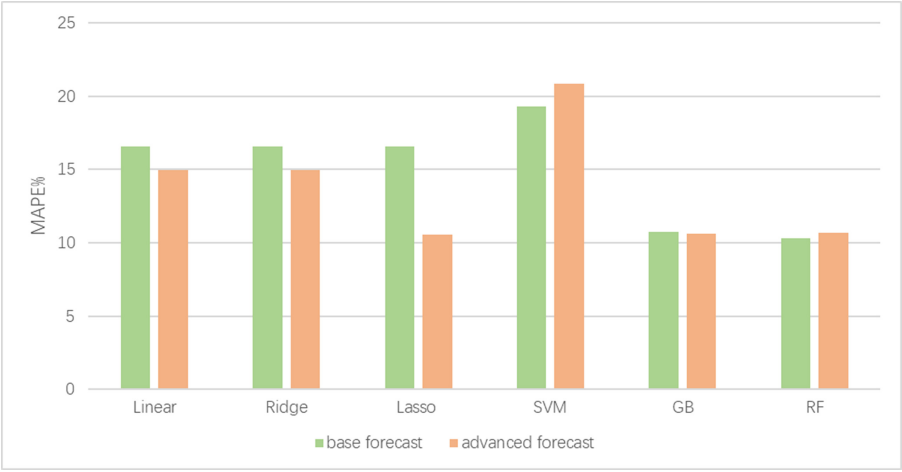

Fig. 2. Mean value of out-of-sample MAPE% in base forecast framework and advanced forecast framework

9.69% and 9.70% for Linear and Ridge respectively. The reason for the larger improvement of Lasso is that Lasso can transform high-dimensional data into a low dimension and preserve the information of high dimension data. When the discount information is included, Lasso can make use of the discount information to make a better forecast, not affected by the increase of features. It is concluded that Lasso is better able to capture the relationship between discount information and sales information and generate more accurate forecasts.

Considering all the above characteristics of models, when the companies have requirements for computing time or the volume of data used is large, an advanced forecast framework with Lasso would be recommended, particularly if they consider incorporating discount information, for its fast computing ability and high accuracy. When the companies require a model that can achieve high accuracy with different kinds of data, the RF and GB would be recommended for their natural broad adaptation and high accuracy.

4 Conclusions

This study has two main contributions. First, the short-term forecast framework proposed can achieve good and robust performance with out-of-sample MAPE ranging from 10.23% to 20.83% in different machine learning models. The three linear models which are linear regression, ridge, and lasso have out-of-sample MAPE ranging from 10.55% to 16.54%. The two ensemble machine learning models which are GB and RF perform best with an out-of-sample MAPE ranging from 10.23% to 10.70%, consistent with their reputation for high accuracy. SVM has the worst performance with a MAPE of about 20% for its poor ability to process high dimension data. Second, considering the discount information improves sales forecast accuracy in linear regression, ridge regression, and lasso regression but doesn't improve the other three models, by the MAPE.

The lasso regression improves largest with 36.19% for it can transform high-dimensional data into a low dimension without losing much information.

References

1. S. J. Canavan. (1997, May 31). Evaluation of Sale Forecasting Methods: A Case Study. Digital Common. https://digitalcommons.njit.edu/theses/998/
2. Y. Li, Y. Yang, K. Zhu, J. Zhang, Clothing sale forecasting by a composite GRU–prophet model with an attention mechanism, IEEE Transactions on Industrial Informatics, vol.17, no.12, 2021, pp.8335-8344. DOI: https://doi.org/10.1109/TII.2021.3057922
3. Y. F. Chen, R. C. Cheng, Single discount or multiple discounts, International Journal of Technology and Human Interaction, vol.15, no.1, 2019, pp.1-14. DOI: https://doi.org/10.4018/ijthi.2019010101
4. R. Fildes, The evaluation of extrapolative forecasting methods, International Journal of Forecasting, vol.8, no.1, 1992, pp. 81-89. DOI: https://doi.org/10.1016/0169-2070(92)90009-x
5. X. Dairul, Z. Shilong, Machine learning model for sales forecasting by using XGBoost, 2021 IEEE International Conference on Consumer Electronics and Computer Engineering, 2021, pp. 480–483. DOI: https://doi.org/10.1109/ICCECE51280.2021.9342304
6. F. Pallonetto, C. Jin, E. Mangina, Forecast electricity demand in commercial building with machine learning models to enable demand response programs, Energy and AI, vol.7, 2021, pp. 100121. DOI: https://doi.org/10.1016/j.egyai.2021.100121
7. H. Du and C. Bo, Sale forecasting method in dynamic environment based on ARMA(1,1), 2011 International Conference on Electric Information and Control Engineering, 2011, pp. 4445–4448. DOI: https://doi.org/10.1109/ICEICE.2011.5777454.
8. J. Zhang and J. Li, Retail commodity sale forecast model based on data mining, 2016 International Conference on Intelligent Networking and Collaborative Systems, 2016, pp. 307–310. DOI: https://doi.org/10.1109/INCoS.2016.42
9. L. Al Hajj Hassan, H. S. Mahmassani, Y. Chen, Reinforcement learning framework for freight demand forecasting to support operational planning decisions, Transportation Research Part E: Logistics and Transportation Review, vol. 137, 2020, p. 101926. DOI: https://doi.org/10.1016/j.tre.2020.101926
10. R. Cui, S. Gallino, A. Moreno, D. J. Zhang, The Operational Value of Social Media Information, Production and Operations Management, vol. 27, no. 10, 2018, pp. 1749–1769. DOI: https://doi.org/10.1111/poms.12707.
11. F. Haselbeck, J. Killinger, K. Menrad, T. Hannus, D. G. Grimm, Machine Learning Outperforms Classical Forecasting on Horticultural Sales Predictions, Machine Learning with Applications, vol. 7, 2021, pp. 100239. DOI: https://doi.org/10.1016/j.mlwa.2021.100239
12. T. Falatouri, F. Darbanian, P. Brandtner, C. Udokwu, Predictive Analytics for Demand Forecasting – A Comparison of SARIMA and LSTM in Retail SCM, Procedia Computer Science, vol. 200, no. 2019, 2022, pp. 993–1003. DOI: https://doi.org/10.1016/j.procs.2022.01.298
13. Y. Ensafi, S. H. Amin, G. Zhang, B. Shah, Time-series forecasting of seasonal items sales using machine learning – A comparative analysis, International Journal of Information Management Data Insights, vol. 2, no. 1, pp. 100058. DOI: https://doi.org/10.1016/j.jjimei.2022.100058
14. L. Breiman, Statistical modeling: the two cultures (with comments and a rejoinder by the author), Statistical Science, vol. 16, no. 3, 2001. DOI: https://doi.org/10.1214/ss/1009213726
15. J. H. Friedman, Stochastic gradient boosting, Computational Statistics and Data Analysis, vol. 38, no. 4, 2002, pp. 367–378. DOI: https://doi.org/10.1016/s0167-9473(01)00065-2

16. X. Yan, Y. Bai, S. C. Fang, J. Luo, A kernel-free quadratic surface support vector machine for semi-supervised learning, Journal of the Operational Research Society, vol. 67, no. 7, 2016, pp. 1001-1011. DOI: https://doi.org/10.1057/jors.2015.89
17. J. Cuesta, C. Matrán, L_p-linear regression, consistency and significative regression in median, Journal of Statistical Planning and Inference, vol. 13, 1986, pp. 15-30. DOI: https://doi.org/10.1016/0378-3758(86)90115-
18. J. Lee, Z. Shi, Z. Gao, On LASSO for predictive regression, Journal of Econometrics, vol. 229, no. 2, 2022, pp. 322-349. DOI: https://doi.org/10.1016/j.jeconom.2021.02.002
19. M. Lee, An iterative method for flattering the ridge in the Ridge regression, Applied Economics Letters, vol. 14, no. 7, 2007, pp. 529-531. DOI: https://doi.org/10.1080/13504850500425840
20. P. M. C. de Boer, C. M. Hafner, Ridge regression revisited, Statistica Neerlandica, vol. 59, no. 4, 2005, pp. 498–505. DOI: https://doi.org/10.1111/j.1467-9574.2005.00304.x