Integration of Survey Data in R Based on Machine Learning

Mattia Spaziani (e-mail: spaziani@istat.it)
Italian National Institute of Statistics (Istat), Rome, Italy

Doriana Frattarola (e-mail: frattarola@istat.it)
Italian National Institute of Statistics (Istat), Rome, Italy

Marcello D'Orazio (e-mail: madorazi@istat.it)
Italian National Institute of Statistics (Istat), Rome, Italy
Food and Agriculture Organization of the United Nations, Rome, Italy

ABSTRACT

This work introduces a relatively new procedure for integrating social survey data through combined use of machine learning techniques and well-known statistical matching methods. The integration is performed with the scope of studying the relationship between variables not jointly observed in the same survey. The results of the new matching procedure seem promising since they are better than those provided by traditional matching methods.

Keywords: *statistical matching, household surveys, supervised classification*
JEL classification: *C40, C80, I31*

1. INTRODUCTION

In National Statistical Offices the data integration is becoming the key step of new statistical production processes aimed at exploiting the already existing data to produce and disseminate a richer set of statistics. It permits to satisfy requests from external users in a timely manner and without increasing the burden on respondents, since no new comprehensive surveys have to be set up.

The integration of data from sample surveys is usually based on *statistical matching*, also known as *data fusion*, that denotes a wide set of methods whose objective is to study the relationship between variables not jointly observed in the same survey (D'Orazio et al., 2006a). In the Italian National Institute of Statistics (Istat) these methods have been extensively applied to integrate the survey on household income with the one on expenditures to get a more comprehensive picture of the living conditions of the Italian households. The experiments investigated a wide set of statistical

matching methods ranging from simple hotdeck to more complex procedures designed to deal with complex sample survey data (Donatiello et al. 2016a, 2016b).

This work illustrates potential use of *machine learning* techniques in the integration of households' survey data. In particular, the proposal reflects the findings of the experiments carried out by D'Orazio (2019a). Since the household surveys collect mainly categorical variables, our application will compare different supervised classification methods, ranging from simple naïve Bayes to boosting-based procedures. The application of machine learning in the statistical matching is allowed by the variety of R packages implementing the machine learning techniques.

The next Section will give an overview of popular statistical matching methods, it will also summarize the work done at Istat to integrate the data from household surveys. Section 3 illustrates the characteristics of the proposed procedure for matching survey data through machine learning. The results of the application of the new procedure to real survey data are presented in Section 4.

2. STATISTICAL MATCHING TO INTEGRATE SOCIAL SURVEY DATA

Let's assume that A and B are two sample surveys referred to the same target population U, that lack of units' identifiers and where the chance of observing the same unit in both is 0. The surveys share a set of X variables (*common variables*), while Y is observed only in A and Z only in B; in practice there is no data to study the relationship between Y and Z. In this setting, statistical matching (SM from now on) attempts to exploit all the available common information, typically a subset of the Xs, to investigate the relationship between Y and Z. In practice, the SM techniques may be used to estimate a parameter of interest (correlation/regression coefficients between Y and Z, contingency table $Y \times Z$, etc.) or to create a *synthetic* ("fused") dataset that, including all variables of interest, serves as basis for further analyses. In this latter case, usually the smaller dataset becomes the reference one (*recipient*) and it is filled in with the missing variable using information "extracted" from the other dataset, i.e. the *donor*. If, for instance, A is the recipient (it includes Y and X variables), then the synthetic dataset is obtained by imputing in it the values of Z using B as the donor.

A number of methods can be used for SM purposes (for a review see D'Orazio et al, 2006a), popular ones are those relying on linear regression models, i.e. it is fitted a regression model on B using Z as the response and

then the imputation of Z in A is based on predictions provided by the model. In alternative, it is possible to resort to nonparametric hotdeck imputation procedures, popular ones are *random* and *nearest neighbor donor* hotdeck. In the first case, the observations in both recipient and donor dataset are grouped in pools of homogeneous units formed by considering values of a subset of the available X variables, then for each recipient unit in a given pool it is chosen at random a donor belonging to the same pool, the value of Z observed on the donor is assigned to the recipient unit. In nearest neighbor donor (NND) hotdeck each recipient unit receives the value of Z observed on the closest donor in B, the distance is calculated on a subset of the Xs. In general, the subset of variables X_M, selected from all the available common variables ($X_M \subseteq X$), that serves for matching purposes is denoted as *matching variables* set; it is usually formed by the "best" predictors of Y and/or Z (for major detail on the selection of the matching variables see D'Orazio et al, 2019)

D'Orazio et al. (2006a) highlight that all the SM methods where the integration is based solely on the matching variables implicitly assume independence between Y and Z conditional on the X_M themselves. Unfortunately, this assumption is seldom valid, unless the X_M include at least a variable highly associated/correlated with Y or Z, i.e. a *proxy* variable. When the conditional independence (CI) assumption is not valid, then the application of SM methods relying on it provides unreliable results as far as the relationship between Y and Z is concerned. In this case, a possible alternative strategy consists in tackling the problem in macro terms by focusing on a target parameter (correlation coefficient ρ_{yz}, cell probabilities $p_{y=j,z=k}$, etc.) for which it is estimated the set of its equally possible values given the available data (D'Orazio et al 2006a, 2006b). In practice, the goal becomes the estimation of the *partial identification regions* for the target parameters. When the "size" of the regions is small there is low "uncertainty" and SM can be performed also under CI assumption; wide regions imply high uncertainty and the SM can only be performed if some auxiliary information can be introduced, so as to reduce the uncertainty itself (cf. D'Orazio et al, 2006b). Assessing the level of uncertainty is not straightforward, a rough measure of uncertainty is suggested in D'Orazio et al (2006b) other approaches are presented in Conti et al (2012) or Zhang (2015).

The package **StatMatch** (D'Orazio, 2015, 2019b) makes easy carrying out SM in the R environment; it implements popular hotdeck SM techniques and includes also methods to integrate data from complex sample surveys; the procedures for assessing the uncertainty are implemented too. The packages helps also in the preliminary SM steps (choice of the matching variables) as well as in evaluating the matching results (see D'Orazio, 2017).

Traditionally, the analysis of economic wellbeing has focused on data on income or consumption expenditures. Unfortunately, the joint collection of income and consumption in a new survey would imply a significant reporting burden on respondents and is not feasible due to budget constraints. In this context, SM represents one of the key instrument to produce statistics on wellbeing by combining data from different surveys. For this reason, in 2012 Istat started a project aimed at integrating data observed in the EU Statistics on Income and Living Condition (EU-SILC) with those from the Household Budget Survey (HBS). Initial application of SM assumed CI and random hotdeck was applied to match the surveys. Failure of CI assumption led to move the focus on how to better exploit the available information (the Italian SILC 2012, with income reference year 2011, and HBS 2012) so as to improve the matching outputs (Donatiello et al, 2016a). Some of the work done focused on the potential use of additional information in order to make the CI assumption an appropriate model for our variables; in particular, we decided to use the information enclosed in the HBS ad hoc section on income and savings (specific module in the Italian HBS) with the objective of estimating a synthetic income variable. This latter variable, being a strong proxy of household income, was used as one of the matching variables, making therefore the CI assumption a reasonable hypothesis (Donatiello et al, 2016a).

Another issue, directly linked to the nature of official survey data, comes from the difficulty to maintain the assumption of independent and identically distributed observations when the units (households) observed in the surveys have been selected through a complex sampling design (e.g. stratified two-stage sampling). In applying the SM at Istat, the complex sampling design issue was addressed through a *weights' calibration* approach, as suggested by Renssen (1998), particularly suited to handle categorical variables. Furthermore, we explored the extension of Renssen's procedure to handle continuous target variables; a two-steps procedure was introduced in line with a mixed approach to SM (Donatiello et al, 2016b).

In 2017 the Italian SILC was extended to include a "Consumption & Wealth" module to collect variables on consumption for food-at-home, public and private transport, regular savings and food-outside-home expenditures. In fact, some analyses showed that food, housing and transport expenditures are very good predictors of the total consumption. This additional SILC module should allow the estimation of a "synthetic" consumption variable that, used as a matching variable in SM, should provide a strong proxy of the consumption and make therefore CI assumption holding.

In recent years emerged the need for a multidimensional approach to household economic wellbeing which takes into account income, consumption

expenditures and also wealth. To this aim, recently Istat started a collaboration with the Bank of Italy to explore possibility of extending SM to exploit their data on wealth collected in the survey on Household Income and Wealth (SHIW).

3. USE OF MACHINE LEARNING TECHNIQUES IN STATISTICAL MATCHING

Machine learning (ML, from now on) involves a wide set of algorithmic-based techniques that "learn from the data" (Hastie et al, 2009). Nowadays they are very popular in marketing, finance, and other domains, also because allow analysis of large data sources, with many variables and observations. ML methods may deal with classification, regression and clustering.

Use of ML in SM is not new, D'Orazio et al (2006a) suggest using classification or regression trees to select of the matching variables (X_M). More in general, it is easy to show that some popular SM hotdeck techniques are particular implementations of the *k*-nearest neighbors (*k*-NN) approach.

A "direct" use of ML in SM is proposed in D'Orazio (2019a) with the objective of creating a synthetic dataset or assessing the uncertainty. In the first case, the sample *B* is used to train a prediction model that is applied to the recipient dataset, *A*, to predict in it the values of *Z* (direct or randomization-based predictions). The results, however, are not particularly satisfactory since the distribution of the imputed *Z* is often unreliable. The second suggestion consists in assessing the uncertainty conditional on the predictions of *Y* and *Z* provided by the ML methods (instead of conditional on X_M). The results in this latter case are more promising since a general reduction of the uncertainty is achieved, avoiding also the major disadvantages related to the direct conditioning on the X_M (e.g. the estimation of relative frequencies in large sparse contingency tables; see D'Orazio et al, 2017 and 2019).

The findings of D'Orazio (2019a) provide the basis for introducing a two-steps SM procedure that goes a step further in the adoption of other modern classification or regression techniques in SM. In particular, the ML methods are used in the first step of the procedure, while the second step consists in the application of a traditional SM hotdeck technique; the whole procedure resembles a mixed approach to SM (D'Orazio et al, 2006a) where the ML methods are fitted in place of traditional regression models. The procedure is articulated as follows:

Step 1) use ML to predict Y and Z in both A and B. In particular (1a) train in A a prediction model for Y and apply it to both A and B to impute in them the predicted values of Y ($\boldsymbol{Y}^*$); similarly, (1b) train in B a prediction model for Z and apply it to impute in both A and B the corresponding predictions ($\boldsymbol{Z}^*$).

Step 2) Set A as recipient and B as donor and apply SM hotdeck (NND of random) to impute the observed values of Z from B to A. In practice, the matching variables of the SM hotdeck are the two predicted variables, $\boldsymbol{Y}^*$ and $\boldsymbol{Z}^*$, obtained in step (1).

A wide variety of ML techniques can be used in step (1). The choice depends on the type of the target variables; if the target variable (response) Y (or Z, or both) is categorical, then the search should be restricted to *supervised classification* methods. On the contrary *supervised regression* methods should be considered when the response variable is quantitative.

The relatively new procedure introduced here presents various advantages when compared to "standard" hotdeck methods. It permits to skip analyses for selecting the matching variables ($\boldsymbol{X}_M$) (see e.g. D'Orazio et al, 2019), needed in SM hotdeck to form the pools of donors or to calculate the distance between units. It is a critical task that, unfortunately, may be a source of difficulties in applying the SM. For instance, in the random hotdeck the crossing of too many Xs may determine empty pools of donors. Moreover, choosing too many matching variables may give an imputed variable with unreliable marginal distribution. When the matching variables include both categorical and continuous Xs, the calculation of the distance is not straightforward (the Gower's distance is usually the privileged solution); this problem is not encountered when applying ML methods that, in most of the cases, can handle mixed-type predictors. Another advantage of ML techniques resides in their ability to detect complex interactions between the response variable and the potential predictors.

Tuning ML algorithms requires however a non-negligible effort, mainly when the set of the Xs is quite large. The required computational effort is heavier when compared to SM hotdeck, where the major IT problem is rather the storage of large matrices of recipient-donor distances.

4. APPLICATION OF THE NEW PROCEDURE TO INTEGRATE REAL SURVEY DATA

This section explores the behavior of the two-steps SM procedure introduced in the previous Section when applied to match data from real social surveys. Section 4.1 describes the survey data being considered while ML methods selected as candidate for step (1) are illustrated in Section 4.2. The results are summarized in Section 4.3.

4.1 The data

In Italy the Household Budget Survey (HBS) and EU Statistics on Income and Living Condition (SILC) survey are both carried out by Istat; the surveys cover the same target population (private households) and the samples are selected by means of a stratified two-stage sampling design.

The focus of HBS is the household (HH) consumption expenditures on goods and services, while SILC collects data on income, poverty, social exclusion and living conditions. The HBS for reference year 2011 provides data on 22,933 responding HHs, while SILC 2012 (income reference year 2011) consists of 18,487 responding HHs. The two surveys show a large number of common variables whose quality and coherence are quite good; the ones selected in this work are a small subset of eight variables, measured at household or personal level (referring to the householder) (see Table 1).

In our matching application, the HBS represents the donor dataset while SILC is the recipient; the HH income (categorized in 7 classes) is the *Y* target variable in SILC; the HH overall expenditures (categorized in 11 classes) plays the role of *Z* in HBS; the objective is to to impute the consumption expenditures classes (*Z*) in SILC (*A*).

Selected set of common variables in HBS and SILC survey

Table 1

Unit	Variable	Categories
Household reference person	Gender	Male, Female
	Marital status	Single, Married, Divorced, Widowed
	Age in classes	<16, 16-24, 25-44, 45-64, >64
	Educational level	No education, Primary, Lower Secondary, Post-Secondary, Upper Secondary, Tertiary
Household level	Geographic macro-areas	Nord-West, Nord-East, Center, South, Island
	Number of employed people	0, 1, 2, >2
	Number of income earners	0, 1, 2, >2
	Number of durable goods	<5, 5, 6, 7, 8

4.2 Chosen machine learning procedures

Since *Y* and *Z* are both categorical, the choice has fallen on a set of popular supervised classification methods which are implemented in R: (i) naïve Bayes classifier; (ii) C5.0; (iii) random forest; (iv) adaptive boosting; and (v) extreme gradient boosting.

The Naïve Bayes (NB) algorithm is based on a simple application of the Bayes theorem for classification purposes (see e.g. Kuhn and Johnson, 2013). In particular, NB assumes that each predictor variable *X* is independent from any other available predictor once conditioning on the target response variable. This "naïve", because fairly unrealistic, assumption provides a significant reduction in the computational complexity. The predicted category is usually the one with the maximum estimated posterior probability.

C5.0 is an improved version of C4.5 algorithm; the objective is to build the tree by reducing the overall entropy or to derive a rule-based classifier (for details see Kuhn and Johnson, 2013). A rule set consists in a series of *if-then* conditions that define a unique route to one terminal node for any record The C4.5 model works by partitioning the sample based on the predictor that provides the highest information gain. Compared to C4.5, C5.0 has some additional features (see Kuhn and Johnson, 2013) and allows also for boosting, a powerful method to increase the accuracy of the classification, as explained later on.

Random forests (Breiman, 2001) focus on ensembles of decision trees for regression or classification problems. Two types of randomness are embedded into the trees: at first, each tree is built from a sample of the starting observations drawn with replacement (bootstrap samples); secondly, at each node, the tree is growth considering the best split among a random subset of input predictors. In practice, a series of independent trees is grown and, in classification problems, the final predicted class it the one getting the majority of votes.

Adaptive boosting (AdaBoost) is a very popular boosting algorithm; it combines ensembles of "weak" classifiers together to improve the classification performance (see Hastie et al, 2009). AdaBoost technique trains the models sequentially by adjusting the weights assigned to the observations: misclassified units in the previous iteration of the classifier have their weights increased. At each iteration the algorithm shifts its focus on the more challenging regions to classify. The predictions are the weighted majority of the fitted trees, where higher weights are assigned to more accurate classifiers in the sequence. The original algorithm was developed for binary response variables; in this work it is considered the multi-class AdaBoost Breiman's extension, where the single weak classifier is a classification tree.

Extreme Gradient Boosting (XGBoost) is an advanced implementation of a gradient boosted tree algorithm (Chen and Guestrin, 2016). Like AdaBoost, models are added sequentially and the method generalizes them through the optimization of an arbitrary differentiable loss function. In particular, by computing second-order gradients of the loss function for different base learners, this method gathers more information about the direction of gradients and how to get to the minimum of the loss function. XGBoost also includes a regularization term that penalizes the complexity of the model, thus preventing overfitting.

The training of the selected supervised classification algorithms on the survey data has been done by means of the R package **caret** (Kuhn et al, 2019), a very comprehensive tool for building machine learning models in R. **caret** provides a common interface to a wide variety of machine learning algorithms and can handle almost every part of the model building process, like cross-validation and parameter tuning.

4.3 Main results

The two-steps procedure is applied to impute the classes of the HH consumption (*Z*) in SILC (*A*) through random hotdeck (i.e. donor chosen at random in donation pools), where the pools of donors are obtained by crossing the predictions, Y^* and Z^*, provided by the chosen ML methods. This two-step procedure is compared with the outcomes of a traditional random hotdeck procedure where the pool of donors are formed by crossing directly two or three of the available *X*s. In R, the random hotdeck step is performed using facilities of **StatMatch** package (D'Orazio, 2015, 2019b) while application of ML algorithms is done mainly via the **caret** package (Kuhn et al, 2019).

The Table 2 summarizes the main results of the different matching exercises. In particular, it is compared how the SM behaves in preserving the marginal distribution of *Z*, as well as the joint *X*-*Z* distribution in the synthetic dataset. To this purpose, the Hellinger's distance between the distribution estimated from the synthetic dataset and the corresponding one observed in the donor dataset (considered as the reference) is calculated. Just for the two-steps procedure, it is reported the prediction error estimated on the training dataset (*A* for *Y* and *B* for *Z*). Finally, the synthetic dataset is used to estimate the association between *Z* (imputed) and observed *Y*, by means of the Cramer's *V*.

Result of statistical matching of HBS and SILC

Table 2

Matching method	Donation classes formed by crossing	Prediction error		Preservation of distributions in the synthetic dataset		Estimated association $Y \times Z$
		Y	*Z*	*Z*	$X \times Z$	(Cramer's *V*)
Traditional	Best 2 *X*s			0.0359	0.1765	0.1666
	Best 3 *X*s			0.0263	0.1785	0.1684
Two-steps	Pred. of naïve Bayes	68.7%	85.3%	0.0267	0.2028	0.1660
	Pred. of C5.0	43.0%	56.4%	0.0278	0.1907	0.1823
	Pred. of randomForest	38.6%	51.7%	0.0256	0.1891	0.1830
	Pred. of AdaBoost	35.9%	49.3%	0.0240	0.1910	0.1882
	Pred. of XGBoost	36.3%	49.7%	0.0270	0.1898	0.1926

The results in Table 2 show that the available *X*s permit to predict better *Y* (the HH income in classes) than *Z* (the HH consumption in classes). The two-steps procedure based on naïve Bayes performs poorly if compared to the remaining ones; the best results are observed with boosting (AdaBoost and XGBoost). It is worth noting that here we are not just interested in prediction accuracy, but rather in creating a proxy that catches as better as possible the prediction power of the available *X*s. Obviously, the smaller is the prediction error the higher are the benefits in using the predicted variable as an input for the marching.

The distribution of the imputed *Z* is fairly close to the one observed in the donor in almost all the cases, a slightly higher value of the Hellinger's distance is observed with traditional random hotdeck when the pool of donors are created by crossing the best two predictors, chosen from the available *X*s. When looking at the preservation of the joint distribution *X*-*Z*, it seems that the traditional hotdeck performs slightly better than the two-steps procedure. In all the cases the estimated values of the Hellinger's distance bring to conclude that all the matching attempts provide an estimate of the joint *X*-*Z* distribution that is not coherent with the reference one (as observed on *B*).

Finally, the Cramer's *V* provides an estimate of the *Y*-*Z* association calculated on the synthetic dataset. The values are obviously smaller than expectations, mostly because, having performed the matching under the CI assumption, the estimated association basically reflects that explained by the considered matching variables, i.e. only a part of the whole association (it is known that CI is not valid in this case). It is worth noting, however, that the estimates of *V* obtained from the synthetic files provided by the two-steps procedure are slightly higher than those related to the traditional random

hotdeck SM, with exception of naïve Bayes. Again the highest estimates of V are those provided by the boosting-based two-steps procedures which, as expected, seem to behave better in catching the interaction between the response variable and the available predictors.

5. CONCLUSIONS

The procedure presented in this work attempts to improve the results of the SM by exploiting the ability of ML procedures in catching complex interactions between the response and the available predictors. It permits to bypass and overcome SM time-consuming task related to the choice of the matching variables, however, the price to pay is a non-negligible effort in tuning the ML algorithms.

The application of this two-steps procedure to match the data from surveys on the Italian households are promising since it permits to improve results when compared to traditional popular SM techniques. Among the tested supervised classification algorithms, the boosting-based ones proved to perform better than the other ones. However, further investigation is deserved to understand how to manage cases with a large number of X variables, being of mixed type (categorical nominal, categorical ordered, numeric). In addition, the two-steps procedure needs to be extended to deal with cases in which one or both the response variables (Y and Z) are numeric.

References

1. **Breiman, L.**, 2001, *"Random Forests",* Machine Learning, 45, pp. 5-32.
2. **Chen T., Guestrin C.**, 2016, *"XGBoost: A Scalable Tree Boosting System"*, eprint arXiv:1603.02754.
3. **Conti P.L., Marella D., Scanu M.**, 2012, *"Uncertainty analysis in statistical matching"*. Journal of Official Statistics, 28, pp. 69-88.
4. **Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M., Spaziani M.**, 2016a, *"The role of the conditional independence assumption in statistically matching income and consumption"*. Statistical Journal of the IAOS, 32, pp. 667-675.
5. **Donatiello G., D'Orazio M., Frattarola D., Rizzi A., Scanu M, Spaziani M.**, 2016b, *"The statistical matching of EU-SILC and HBS at ISTAT: where do we stand for the production of official statistics"*, DGINS - Conference of the Directors General of the National Statistical Institutes, 26-27 September 2016, Vienna
6. **Hastie T, Tibshirani R, Friedman J.**, 2009, *The Elements of Statistical Learning.* 2nd Edition. Springer, New York.
7. **Kuhn M, Johnson K.**, 2013, *Applied Predictive Modeling*. Springer, New York
8. **Kuhn M.** with contributions from **Wing J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, Mayer Z, Kenkel B, the R Core Team, Benesty M, Lescarbeau R, Ziem A, Scrucca L, Tang Y, Candan C.** and **Hunt T**., 2019, *"caret: Classification and Regression Training"*. R package version 6.0-82. https://CRAN.R-project.org/package=caret

9. **D'Orazio M.**, 2015, *"Integration and imputation of survey data in R: the StatMatch package"*. Romanian Statistical Review, 2/2015, pp. 57-68.
10. **D'Orazio M.**, 2017, *"Statistical Matching and Imputation of Survey Data with StatMatch"*. StatMatch R package Vignette
11. https://cran.r-project.org/web/packages/StatMatch/index.html
12. **D'Orazio M**, 2019a, *"Statistical Learning in Official Statistics: the Case of Statistical Matching"*. Submitted for publication on the Statistical Journal of the IAOS.
13. **D'Orazio M**., 2019b, *StatMatch: Statistical Matching or Data Fusion*. R package version 1.3.0.
14. https://CRAN.R-project.org/package=StatMatch
15. **D'Orazio M, Di Zio M, Scanu M.**, 2006a, *Statistical Matching, Theory and Practice*. Wiley, Chichester
16. **D'Orazio M, Di Zio M, Scanu M**., 2006b, *"Statistical matching for categorical data: displaying uncertainty and using logical constraints"*. Journal of Official Statistics, 22(1): pp. 137–157
17. **D'Orazio M, Di Zio M, Scanu M.**, 2017, *"The use of uncertainty to choose matching variables in statistical matching"*. International Journal of Approximate Reasoning, 90, pp. 433-440.
18. **D'Orazio M, Di Zio M, Scanu M**., 2019, *"Auxiliary variable selection in a statistical matching problem"*. In: Zhang LC, Chambers RL, (eds.). *Analysis of Integrated Data*. CRC Press, Boca Raton, pp. 101-120.
19. **Renssen R.H.**, 1998, *"Use of Statistical Matching Techniques in Calibration Estimation"*. Survey Methodology, 24, pp. 171–183.
20. **Zhang LC.**, 2015, *"On proxy variables and categorical data fusion"*. Journal of Official Statistics, 31, pp. 783-807.