

**2013 Federal Committee on Statistical Methodology (FCSM) Abstract Template**

Please use this Microsoft Word abstract template to prepare your abstract for the 2013 FCSM Research Conference. Send your abstract via email to [fcsm.conf@gmail.com](mailto:fcsm.conf@gmail.com) no later than **April 8, 2013**.

**1. Author Contact Information**

**Primary Author Information**

Please provide the following information for the primary author:

|  |  |
| --- | --- |
| **First Name** | John |
| **Middle Initial** | M |
| **Last name** | Abowd |
| **Organization** | Cornell University |
| **Mailing address** | 261 Ives Hall Cornell University  Ithaca, NY 14853 |
| **Country** | USA |
| **Phone** | +1607-255-8024 |
| **Email** | john.abowd@cornell.edu |
| **Fax** |  |

**Secondary Author(s) Information**

Please provide the full name, affiliation, and email for each co-author, if applicable. Add additional rows if necessary.

|  |  |
| --- | --- |
| **Name and Affiliation** | **Email** |
| Ping Li, Cornell University | pingli@cornell.edu |
|  |  |
|  |  |
|  |  |

**2. Abstract**

Abstracts should include a brief description of the research purpose and focus, data sources used, statistical methods employed, and conclusions reached, as appropriate. The maximum length of abstracts is 300 words. Please use the following formatting instructions when preparing your abstract. See page 2 for an example.

Formatting:

File type: Microsoft Word

Font type: Arial

Font size: 12 pt.

Font style Heading – bold type

Text – regular type

Case: Upper and lower-case letters for all headings and text

Line spacing: Single

Paragraph format: No initial indent, use blank line between paragraphs

**Organized sessions**: Proposals for organized sessions should include 3 papers and a discussant OR 4 papers. Please put all abstracts into one template or send the abstracts together in one email. Include the session title and discussant name (if applicable) in your submission.

**Please mark (X) one:**

\_X\_\_ Contributed paper

\_\_\_\_ Organized session paper (please list other papers included in the proposed session below)

\_\_\_\_ Technical software demonstration

**Keywords: Logistic regression, boosting, race and ethnicity coding, multiple response, multiple label**

**Boosting Algorithms for Edit and Imputation of Multiple-response Variables**

Ping Li (Cornell University), John M. Abowd (Cornell University)

Many practical classification datasets naturally have multiple responses of the form of “check all that apply” variables. Such variables are also known as “multi-label.” A prominent example is the multi-label race variable mandated by the OMB standards for race and ethnicity reporting that is used by the Census Bureau and many other federal agencies. This response is frequently reported as “white alone,” “black or Afro-American alone,” “white and some other race,” *etc.* An additional example occurs in text categorization where a document may belong to several pre-specified topics, such as sports and news. In medical diagnosis, a disease could belong to multiple categories. In general, predicting multiple-response variables is a more challenging task than regular single-response classification. In this paper, we propose a simple statistical framework that generalizes the standard multinomial logistic regression model to accommodate multiple-response variables. We do this without artificially expanding the classification variable to all feasible, as opposed to observed, permutations of the multiple responses. This is the basis for the prediction performance gains we achieve. Our algorithm, implemented based on boosting and trees, is tested on artificial data and on several standard multi-label datasets, including public-use decennial census data. The principle advantage of our approach is that it allows for a more accurate prediction model that can be used as part of the edit and imputation processing for surveys that contain multiple-response variables.