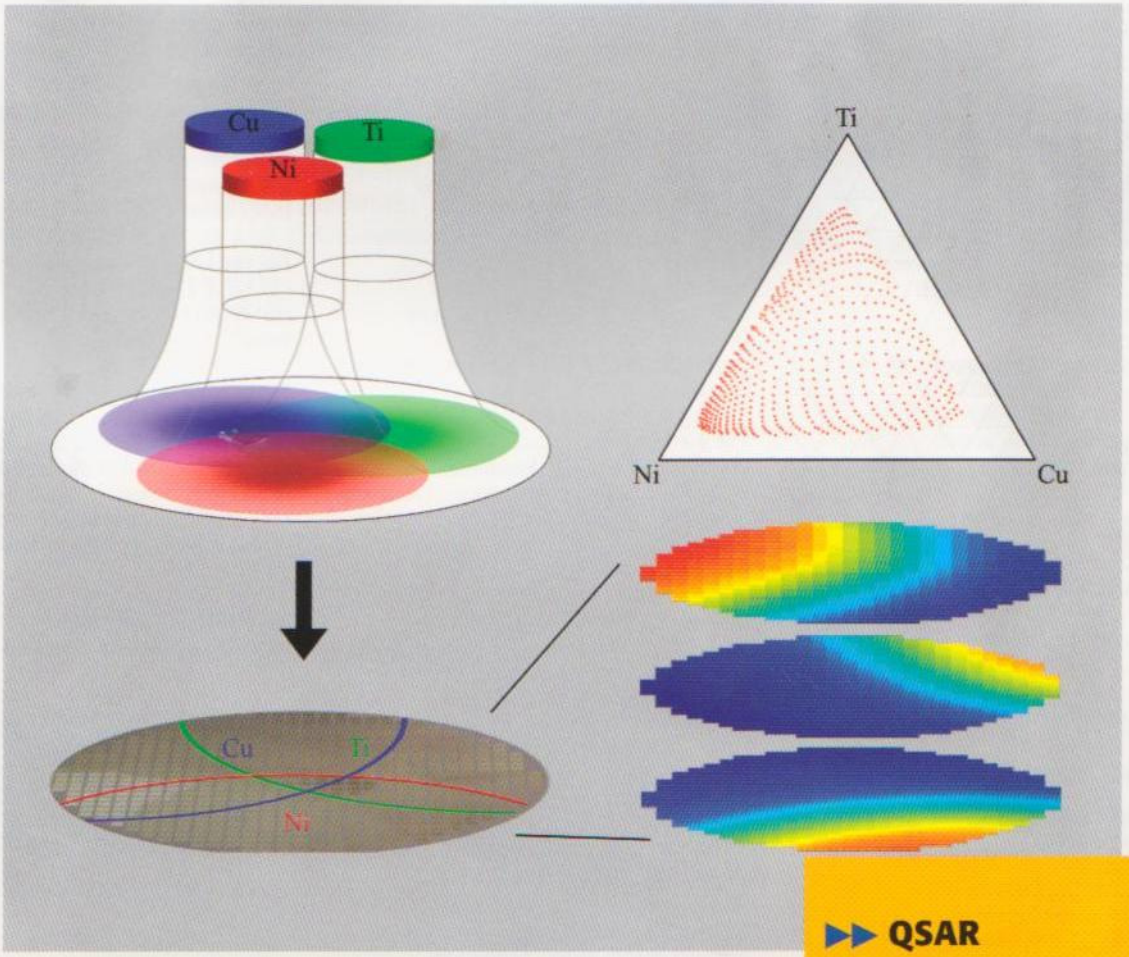
Vol 27 – No 2 | February 2008| QCSSAU 27(2) | 133-268 (2008)

02|08

QCS

QSAR & Combinatorial Science



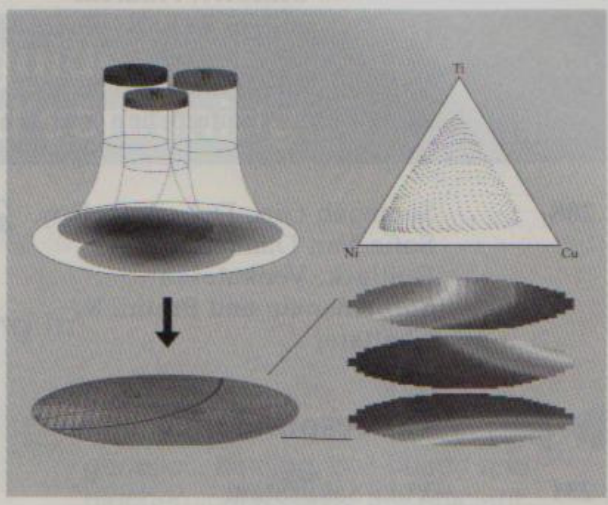
QCS

QSAR & Combinatorial Science

Volume 27, Number 2/2008
February 2008
Pages 133 – 268

Cover Picture:

Co-sputtering of three metallic elements is used to generate composition spread thin films. The gradient thickness profile from each sputtering gun results in continuously changing composition of the deposited ternary film across the wafer which can be directly mapped onto a ternary phase diagram. Various properties are rapidly mapped across the ternary phase diagram. In their Full Paper on pages 171-178. Rossana Dell'Anna, Paolo Lazzeri, Roberto Canteri, Christian J. Long, Jason Hattrick-Simpers, Ichiro Takeuchi, and Mariano Anderle used the supervised principal component technique to investigate the relationship between time of flight secondary ion mass spectrometry (ToF-SIMS) spectra and the composition distribution across a Ni-Ti-Cu composition spread.



Data Analysis in Combinatorial Experiments: Applying Supervised Principal Component Technique to Investigate the Relationship Between ToF-SIMS Spectra and the Composition Distribution of Ternary Metallic Alloy Thin Films

Rossana Dell'Anna[a]*, Paolo Lazzeri[a], Roberto Canteri[a], Christian J. Long[b], Jason Hattrick-Simpers[b], Ichiro Takeuchi[b], Mariano Anderle[a]

* Fondazione Bruno Kessler – irst, Via Sommarive 18, 38050 Povo (Trento), Italy, E-mail: dellanna@itc.it
[b] Department of Materials Science and Engineering, University of Maryland, College Park, MD 20742, USA

Keywords: Composition spreads experiments, Feature selection, Regression models, Supervised principal components, ToF-SIMS spectra

Received: January 30, 2007; Accepted: March 30, 2007

DOI: 10.1002/qsar.200740008

Abstract

We apply a semi-supervised technique called Supervised Principal Component (SPC) to explore the relationship between the composition of a thin film combinatorial library and the peaks of Time-Of-Flight Secondary Ion Mass Spectrometry (ToF-SIMS) spectra acquired from the library. SPC is first used to select a subset of the available multivariate features (the peak intensities of the ToF-SIMS spectra) based on their association with the outcome variable (the elemental concentration of the thin film samples). Next, using only the selected features, SPC creates optimal linear models which map the ToF-SIMS data onto the composition data. The models for the first two of the considered elemental concentrations use only eight of the 55 available ToF-SIMS peaks, making interpretation of the model much simpler than for a model which uses all 55 available peaks. The percentage of explained variance (R^2) in concentration data is in both cases about 0.80. These results are obtained during the model validation phase, performed on test data, which are exclusively used for this purpose. The model for the third considered element did not produce significant results due to the poor variability of the dataset. This work illustrates for the first time that using a multivariate analysis technique, one can establish the relationship between ToF-SIMS measurements and stoichiometric data in a combinatorial experiment. More generally, the described feature selection approach provides an example of how combinatorial experiments can be useful for accelerating the understanding of the chemical–physical behaviors under investigation.

1 Introduction

Combinatorial experiments for the discovery and optimization of new materials [1] generate data that are usually multivariate, as arrays of variables (features) and multiple structural and/or functional outputs are typically associated with a library of compounds. A crucial aim of combinatorial experiments is to exploit these data for developing reliable predictive models which are capable of identifying materials possessing desirable physical properties. By recognizing new patterns in data, data mining strategies could provide new hypothesis on composition-structure–property relationships to be further validated by new experiments or theoretical explanations.

One area of critical importance in data mining is feature selection [2]. Feature selection is crucial in data mining because it helps to filter out both redundant and irrelevant information from a multivariate dataset. The decrease in dimensionality of a multivariate dataset obtained by feature selection not only reduces the computational expense of the algorithm, but can also reduce the risk of designing models which are over-fitted to data. By determining relevant modeling variables, a new insight into the mechanism governing the considered physical behavior is possible. Interpretability, scalability, and, possibly the accuracy of the

Abbreviations: PC, principal component; **PCR,** principal component regression; **SPC,** supervised principal component; **ToF-SIMS,** time-of-flight secondary ion mass spectrometry; **WDS,** wavelength dispersive spectroscopy

