RISK FACTOR VARIABLE SELECTION AND IMPORTANCE RANKING IN THE CONTEXT OF WELL POLLUTION

E. Acar,[†], E. Linder

University of New Hampshire, Durham, NH, USA

[†] E-mail: *efc2@cisunix.unh.edu*

For risk factor identification we examine a data base consisting of (1) pollutant concentrations determined from periodic testing on 1300 public wells over a 10 year period, (2) well characterizations in terms of geology, construction, well operation, (3) number of and distances to potential nearest sources of pollution, such as underground storage tanks, roads, garages, etc. The data base consists of over 300 potential risk factor variables. We perform variable selection for both, the incidence of detectable pollution level, as well as the maximum observed level - 73% of the wells have pollution levels below the detection limit. Because of many missing values we first apply recursive partitioning (classification and regression trees) with crossvalidation based pruning. Secondly we apply stepwise variable selection for semiparametric (spline based) predictive models that include many interactions. Finally we develop an importance ranking of the risk factors. The importance measure is based on a combination of deviance reduction and the order of appearance (for tree models) and on projection pursuit representation (for the regression models). Finally we examine an overall risk factor ranking based on all analyses for the purpose of risk communication to stakeholders.