Application of data clustering to railway delay pattern recognition

K-means clustering is employed to identify recurrent delay patterns on a high traffic railway line north of Copenhagen, Denmark. The clusters identify behavioral patterns in the very large ("big data") data sets generated automatically and continuously by the railway signal system. The results reveal where corrective actions are necessary, showing where recurrent delay patterns take place. Delay profiles and delay-change profiles are generated from timestamps to compare different train runs, and to partition the set of observations into groups of similar elements. K-means clustering can identify and discriminate different patterns affecting the same stations, which is otherwise difficult in previous approaches based on visual inspection. Classical methods of univariate analysis do not reveal these patterns. The demonstrated methodology is scalable and can be applied to any system of transport.