

3. Finally, the degradation profile (DP) and recovery profile (RP), which are based on the weighted sum in Munoz and Dunbar (2015) but do not include a time penalty, describe the summed deviation from a linear degradation and recovery, respectively. 1: where T_0 is the start time of the disruption; T_1 , T_2 and T_3 are the end time of the first, second and third phase, respectively; $f(t)$ is the linear degradation function connecting (T_0, Q_0) and (T_1, Q_{min}) ; $g(t)$ is the linear recovery function connecting (T_2, Q_{min}) and (T_3, Q_0) ; n is the total number of time intervals $t_{i+1} - t_i$, which is equal to the disruption length when taking a step size of one unit; and m and s are the number of equally spaced measurement points in the first and the third phase, respectively. The current DP and RP metrics work well under the assumption of convex/concave behaviour of the performance in degradation/ recovery phases, which is the most commonly observed behaviour in real-life railway disruptions. In the modified BFS algorithm, the start vertex/vertices is/are the disruption location (in case of a timetable point outage) or boundaries (in case of a line blockage), and the end vertices are the second closest decoupling points. To measure both indicators simultaneously, they are combined in a composite performance indicator Q which is calculated as the weighted sum of traffic punctuality (P/R) and traffic intensity (R/S):

3.3. The BFS algorithm starts from a source node, referred to as the start vertex, and visits any adjacent, unvisited vertices until none are left. If no train operations would be possible to and from Apd, this would affect the surrounding lines significantly, and the first decoupling points would be Amf, Dv and Zp. Instead, if limited train operations to and from Apd would be possible, then Apd itself would be the first decoupling point, and Amf, Dv and Zp would be the second decoupling points. The formal procedure of the modified BFS algorithm for a single start vertex is included in Appendix A. For more details on the algorithm, the interested reader may refer to Knoester (2021). Therefore, we also consider the traffic punctuality, which is defined as the proportion of punctual train activities (i.e. with a delay less than three minutes, which is a threshold used at the Dutch railways) relative to the realized train activities in a given time period. Apart from plotting individual resilience curves, one could also identify different shapes of resilience curves and calculate the mean and median resilience curve per disruption type. Disruption management in the Netherlands provides a theoretical foundation for determining which area to consider, based on the concepts of decoupling points and impact areas. 3 (a), line Amf–Apd represents a simple, open line, and thus, two start vertices would suffice (being Amf and Apd) which would also be marked as the first decoupling points. Therefore, such system is deemed more resilient, because the gradual degradation gives traffic controllers more time to respond, hence the smaller DP. Similar holds for recovery profile (RP).

Methodology A data-driven methodology is developed in this study to ex-post evaluate the resilience of railway networks given a large number of disruptions of varying types. The methodology is generic and can work for different performance indicators, resilience metrics and categorizations of disruption types. Performance measurements at each time instant are stored in a dataframe, which serves as the basis for calculating the resilience metrics. The metrics are stored in a separate dataframe, which serves as the basis for conducting statistical analyses. The resilience curves, which can be drawn directly from the performance measurements, represent the graphical output of the resilience quantification. We first introduce the boundaries of a disruption, which could be the decoupling points closest to the disruption location, but could also be the timetable points marking a smaller area in which

there is no traffic operation, depending on the magnitude of the disruption. The original BFS algorithm, introduced by Moore (1959), is a type of graph search that is used to traverse a graph and find each node or vertex in the graph. The degradation time (DT), response time (RST) and recovery time (RCT) describe the duration of the first, second and third resilience phase, respectively. The seven resilience metrics are defined as follows, using the notation indicated in Fig. It is structured in the resilience quantification framework presented in Fig. The framework has been divided into three parts (i.e. input, processing and output) and works as follows. The resilience metrics and the test statistics represent the numerical output of the resilience quantification. To determine the impact area of each disruption given the disruption location, a modified breadth first search (BFS) algorithm is developed. For example, in case of a line blockage between Amersfoort (Amf) and Apeldoorn (Apd) as in Fig. Calculating performance: In order to calculate performance, it is necessary to specify a performance indicator which measures the level of operations at a given time instant. The statistical analyses are used to identify differences among disruptions of varying types. Determining the impact area: A question arising before calculating the performance during a disruption is which area to consider in this calculation. Since disruption effects can easily spread through the network, it is preferred to study a larger area than just the disrupted line or timetable point. A decoupling point is defined as a timetable point where trains are allowed to start or end their route in case of a disruption. An additional benefit of using impact areas is that it limits possible interactions with disruptions in other parts of the network, both in our quantification approach and in practice. For our railway network context, several new extensions are required on the basic BFS algorithm. Fifth, the algorithm needs to handle more than one start vertex, since a single start vertex only occurs for a timetable point outage. A common approach is to express railway system performance in terms of the traffic level (e.g. Ghaemi et al., 2017), which we refer to as the traffic intensity. Traffic intensity is defined as the proportion of realized train activities relative to the scheduled train activities in a given time period. While such a definition for the traffic intensity indicator accounts for cancellations, it does not account for delays. The maximum impact (MI) describes the vertical distance between target performance (Q_0) and minimum performance (Q_{min}). For all metrics, a higher positive value indicates a stronger disruptive effect, and thus, a less resilient network. Looking at the degradation profile (DP), in case that it degrades gradually and takes a long time to reach minimum performance, the DP value will be smaller than when performance drops quickly and then keeps degrading slowly. The quantification starts with the collection of traffic realization data, disruption log data and network data from a database. Disruption log data refer to information about the disruption, such as the reported start and end time, location, cause, etc. If the studied area is too small, then disruption effects further away from the disruption location could be overlooked, but if the studied area is too large, the impact of the disruption becomes less visible. The first impact area is bounded by the first intercity decoupling points from the disruption location; the second impact area is bounded by the next closest intercity decoupling points from the first ones; and the third impact area is bounded by the next closest intercity decoupling points from the second ones. Since cancellations in the third impact area are in principle not allowed, disruption effects are mostly contained within the first and second impact area. Thus, the first and second impact area are identified as the appropriate area to consider for the resilience quantification. Third, the

algorithm must account for the fact that not all vertices may appear in the same path in opposite directions. Fourth, the algorithm must allow a vertex to have more than one parent (i.e. the previous vertex in the specific path), and to visit this vertex when it has already been visited from another parent. Instead, a line blockage will have two, three or even four start vertices, depending on the network layout. Alternatively, assuming a disrupted station in Apd as in Fig. Calculating the resilience metrics: The shape of the resilience curve is described by seven resilience metrics. Performance loss (PL) describes the area enclosed by target performance and the resilience curve. However, due to a possible complex performance variability during some specific disruptions (e.g. a sudden drop in performance during recovery phase), DP or RP could reach close to zero values. Traffic realization data includes at least the plan time, realization time and location of train activities. Network data specify how each timetable point in the network is connected to neighboring timetable points. Using these data, the evolution of performance over time can be calculated for each disruption in a disruption-specific impact area. First, the algorithm terminates its search along a branch of the network when a second decoupling point is reached in the specific branch. 3 (b), one start vertex would suffice (being Apd). Together, the performance indicators account for cancellations as well as delays. The first five metrics can only have nonnegative values, whereas the last two metrics can have positive values as well as negative values. The size of the area depends on the location in the network, but also on the type of impact, as illustrated in Fig. The new extensions are set as follows. The area covered in between represents the impact area. Second, the algorithm cannot follow a path that is not driven by any train. 2.3. 1.3.3.2.