steer davies gleave

Route performance measurement

Report

March 2017

Office of Rail and Road

Our ref: 23095701

Client ref: ORR CT 16-81

Route performance measurement

Office of Rail and Road

Report
March 2017

Our ref: 23095701
Client ref: ORR CT 16-81

Prepared by:

Steer Davies Gleave
28-32 Upper Ground
London SE1 9PD

+44 20 7910 5000
www.steerdaviesgleave.com

Prepared for:

Office of Rail and Road
One Kemble Street
London
WC2B 4AN

Contents

Figures

Tables

Executive summary

Introduction

ORR is seeking to develop a performance metric that can be used in benchmarking the contribution of Network Rail's routes to train performance. The metric must enable comparisons of performance both between routes and over time within a single route. Steer Davies Gleave was commissioned in February 2017 to support the ORR in specifying potential metrics, and this report sets out the findings of our work.

Objectives

At the time of writing, the ORR is still considering whether to adopt a single metric to be used in setting formal regulatory targets for the routes for CP6, or to introduce one or more KPIs to support performance monitoring more generally and provide a framework for challenging routes over the potential for improvement. For the purposes of this work, we have assumed that the metric will be used as a basis for setting formal regulatory targets for the routes, since this requires more thorough consideration of how to ensure meaningful comparability between them. However, we have also considered how the main metric might be supplemented with others supporting a less formal approach to performance monitoring.

If the metric is to serve the purpose described above, it must meet a number of specific objectives. These have been discussed and agreed with the ORR and are set out below:

i. **Comparability** - the metric must enable meaningful comparison of performance between routes, recognising that they differ substantially in terms of their geography, scale and technical characteristics as well as the volume and type of train services that use them.
ii. **Ease of calculation** - it must be straightforward to calculate using data already collected and systems already in place, or that could be supplemented at limited additional cost.
iii. **Consistency** - it must be correlated with other measures used for monitoring performance, in particular those specified in the route scorecards as well as those on which Schedule 8 of the track access agreements is based.
iv. **Customer focus** - it must reflect the experience of the passenger and freight customer.
v. **Credibility** - it must be acceptable to, and capable of being understood by, relevant stakeholders, including route directors, train operators and customer organisations.

In practice, there is some tension between these objectives, and the chosen metric is likely to represent a compromise, meeting some fully but others only partially. For the purposes of appraising the metrics specified and reviewed in this report, we have prioritised the objectives according to the order in which they appear above.

Issues in metric design

Route contribution to performance

It is important to consider the role that the routes have in delivering performance since this must inform any understanding of the performance for which they are accountable. From a performance perspective, this role can be defined as follows:

- **Asset stewardship** – this involves developing and implementing an asset management strategy with a view to avoiding asset failures causing disruption.
- **Timetable planning** – supporting Network Rail's system operator function in developing Train Planning Rules governing the planning of scheduled services.

- **Control according to the timetable** – the route is also responsible for making day-to-day operational decisions that enable train services to run in accordance with the timetable.
- **Management of incidents occurring on the route** – when incidents do occur, they must be managed in order to minimise the impact on train services and their customers.
- **Mitigation of delay imported from other routes** – delay must also be mitigated, which typically requires effective communication between routes as well as with train operators.

Capturing the passenger experience

We have also considered how best to capture the passenger experience using the lateness or delay metrics already reported and supporting the performance incentive mechanism in Schedule 8 of the track access agreements. Lateness is arguably a better measure of success than delay, since it reflects the impact of sub-threshold minutes that are not captured by the delay attribution process as well as correlating more closely with the experience of passengers and users of freight services. However, there is currently no basis for allocating lateness between routes, since the Monitoring Points used to capture lateness do not align with route boundaries. In any event, delay is generally highly correlated with lateness, which means that any metric based on delay will tend to reflect the passenger experience.

Ensuring comparability

Routes differ markedly in their scale and technical characteristics as well as the types of train service that operate on them. This means that in order to compare performance by route, it is necessary to normalise the measure using an appropriate scaling or other factor. In principle, a number of normalisation factors are available:

- **Measures of the scale of the infrastructure:** these are measures of the physical size of the infrastructure, such as route km or track km, which give a broad indication of both the scale of the route's responsibilities and the potential for service disruption.
- **Measures of the level of service activity:** it is also possible to normalise a route performance measure using a factor that directly captures the level of service activity, such as train km or vehicle km. In broad terms, the greater the number of train km operated, the greater the potential for delay and the greater the total delay is likely to be.
- **Measures of the level of passenger traffic:** route performance can also be expressed in terms of delay per passenger journey. Such a measure directly captures the passenger experience, particularly when combined with a passenger-weighted measure of delay.
- **Measures of the level of service intensity:** it is also possible to combine normalisation factors, for example by dividing train km by track km to give a measure of service intensity (reflecting the volume of traffic in relation to the scale of the infrastructure). This approach to normalisation controls for the greater potential for secondary delay to propagate across services on more intensively used parts of the network.

We have investigated the use of both service activity (train km) and service intensity (train km per track km) as a basis for normalisation. In general, we consider service intensity to be the more appropriate basis, at least in the case of metrics that capture both primary and secondary delay, as it takes account of more information about the characteristics of a route.

Specification and evaluation of metrics

We have specified seven metrics for investigation, taking account of the objectives and design issues discussed above. These are set out, together with a summary of the results of their appraisal against the objectives, in the table below. Note that we discuss a wider range of variants on the metrics set out below in the main report (in particular, investigating the impact of using different normalisation factors), but the table gives an indication of the strengths and weaknesses of the main measures.

Metric		Normalisation factor	Summary appraisal
M1: Minutes Delay (MD)	Total Network Rail-caused MD attributed to a route	Service intensity (train km/track km)	• Fails to take account of import and export of delay across boundaries. • Risk of stakeholder challenge.
M2: Passenger-weighted MD	M1 is modified by weighting MD by passenger journeys	As for M1	• Route has limited influence over passenger journeys by service. • Risk of stakeholder challenge.
M3: Performance Minutes	Standard Schedule 8 metric, calculated for each route	Normalised according to standard Schedule 8 calculation	• Normalisation does not take account of key route characteristics. • Cannot be reliably calculated with existing Network Rail systems.
M4: Primary MD	Total primary Network Rail-caused MD attributed to a route	Service activity (train km)	• Allows for meaningful normalisation and comparison as import and export of primary delay is minimal. • Does not reflect passenger experience.
M4+: Primary + secondary MD	Total Network Rail-caused MD (caused and suffered on a given route)	As for M1	• Allows for meaningful normalisation and comparison as imported and exported delay is excluded. • More fully reflects passenger experience than M4.
M5: Secondary MD	Total level of secondary delay incurred on the route, regardless of cause	As for M1	• Inclusion of all secondary delay regardless of cause is useful but reduces comparability (although it could be modified). • Risk of stakeholder challenge.
M6: Deemed Minutes Lateness	Standard Schedule 8 metric, calculated for each route	Normalised according to standard Schedule 8 calculation	• Useful as a supplementary metric. • Cannot be reliably calculated with existing Network Rail systems.

Scores well against objectives			Scores poorly against objectives

Recommendation

In the light of these results, we recommend that the primary metric used for comparing route level performance should be defined as:

M4+: The sum of Network Rail-caused primary and secondary delay caused by, and suffered on, a given route, normalised using train km per track km.

We consider that it should be supplemented by the following additional metrics, reviewed on a less formal basis but nevertheless used to inform ongoing review and challenge of route level performance as part of a broader regulatory engagement process:

- *M4: The sum of Network Rail-caused primary delay caused by, and suffered on, a given route, normalised using train km;*
- *M5: The sum of all secondary delay suffered on a route, regardless of cause, normalised using train km per track km; and*
- *(At a future date following the necessary changes to existing systems) M6: Average deemed minutes attributable to a route, derived using parameters and metrics already used in the calculation of Schedule 8 payment.*

2 Introduction

Background

2.1 The reorganisation of Network Rail in November 2015 cemented an ongoing process of route devolution, providing route directors with greater discretion over operational decision-making and facilitating comparison between different aspects of route performance. As the Office of Rail and Road (ORR) has noted, this should support more effective regulation, both by improving the availability of route-level information and by strengthening the incentives on routes to out-perform each other[1]. At the same time, individual routes will have greater freedom to focus on the requirements of their customers, unconstrained by output decisions taken at the centre.

2.2 There is some potential for these developments to result in tensions within the regulatory framework, particularly in the area of operational performance. On the one hand, a combination of greater discretion and more focus on individual passenger and freight operator requirements will tend to result in a multiplicity of metrics and targets, as can be seen from an examination of the route scorecards already prepared for Control Period (CP) 5. On the other, benchmarking route performance, by definition, requires a common metric or set of metrics that can capture each route's contribution to performance over a given period of time.

2.3 Against this background, the ORR is seeking to develop a performance metric that can be used in benchmarking the contribution of Network Rail's routes to train performance. The metric must enable comparisons of performance both between routes and over time within a single route. It will be used in order to monitor and manage performance during CP6, and will be reported alongside other, route-specific metrics being developed for the new route scorecards to be agreed between routes and their respective customers. It must work in tandem with these bespoke metrics and with the metrics and incentives within Schedule 8 of the track access agreements.

Purpose of report

2.4 Steer Davies Gleave was commissioned in February 2017 to support the ORR in specifying and reviewing potential metrics for comparing route-based performance, and this report sets out the findings of our work. More specifically, it describes a number of metrics and illustrates their application in comparing each of Network Rail's routes before appraising them against a series of agreed objectives. It also provides specific recommendations for the monitoring of route-based performance in CP6, including a recommendation on the most appropriate metric.

[1] PR18 working paper 1: implementing route level regulation, ORR June 2016.

2.5 In line with our terms of reference, we have focused largely on metrics capturing Network Rail's contribution to the performance of passenger services. However, a number of the metrics considered cover all services, and could be further disaggregated to differentiate between passenger and freight.

2.6 The report has been informed by discussions with both ORR and Network Rail covering the specification of suitable metrics, data availability and issues concerning the practicality of different calculations. However, at this stage we have not undertaken any wider consultation among industry stakeholders.

Organisation of report

2.7 The remainder of the report is organised as follows:

- Chapter 3 sets out a series of objectives for the route performance metric;
- In Chapter 4, we discuss of a number of issues that need to be taken into account in the design of the metric if the objectives are to be met;
- Chapter 5 provides a specification of the six metrics and a brief discussion of their strengths and limitations;
- Chapter 6 illustrates the application of the metrics in comparing route-level performance, and discusses their relationship with other measures of performance; and
- In Chapter 7, we report the results of a qualitative appraisal of the objectives against the objectives described in Chapter 3, and set out our recommendations for monitoring route level performance in CP6.

3 Objectives

Purpose of the metric

3.1 As indicated in the previous chapter, the metric is intended to enable comparisons of operational performance between routes and calculation of changes in performance on individual routes over time. However, at the time of writing, the ORR is still considering whether to adopt a single metric to be used in setting formal regulatory targets for the routes for CP6, or to introduce one or more KPIs to support performance monitoring and provide a framework to support the regulator and stakeholders in challenging routes over the potential for improvement. While in both cases the metric will be important in holding the routes to account for the performance that they deliver, the way in which it is used within the regulatory framework has important implications for its design.

3.2 In particular, if the aim is to introduce a metric representing a key regulated output, together with the potential for regulatory sanctions in the event that targets are not met, it is necessary to define a single measure that takes full account of factors affecting route level performance that are beyond a route's ability to manage. Alternatively, the introduction of one or more KPIs as a basis for challenging route directors and informing discussions about performance would arguably require a less rigorous specification. Under this approach, there would be less focus on meeting a specific target, and routes would be held to account through monitoring of a number of measures, all potentially subject to limitations but taken together providing a comprehensive picture of performance.

3.3 For the purposes of this work, we have assumed that the metric will be used as a basis for setting formal regulatory targets for the routes, since this requires more thorough consideration of how to ensure meaningful comparability between them through its specification. However, we have also considered how the main metric might be supplemented with others supporting a less formal approach to performance monitoring.

Specific objectives

3.4 If the metric is to serve the purpose described above, it must meet a number of specific objectives, each informing different aspects of its design. These have been discussed with the ORR and are set out below:

i. **Comparability** - the metric must enable meaningful comparison of performance between routes, recognising that they differ substantially in terms of their geography, scale and technical characteristics as well as the volume and type of train services that use them. While it will not be possible to control for all relevant factors in the specification of a single metric, some normalisation will be needed to ensure that routes can reasonably be held accountable for the relative levels of performance indicated.

ii. **Ease of calculation** - the metric must be straightforward to calculate using data already collected and systems already in place, or that could be supplemented at limited additional cost. This will ensure that monitoring of the metric can be introduced in CP6. At the same time, the specification should be robust to potential changes in data collection and supporting systems, for example changes in route boundaries, the definition of sub-threshold delay, the introduction of new monitoring or recording points, or modifications to key systems such as TRUST and PEARS.

iii. **Consistency** - the metric must be correlated with other measures used for monitoring performance, in particular those specified in the route scorecards (which currently vary by train operator but include measures such as right time performance, the Public Performance Measure (PPM) and Cancellations and Significant Lateness (CaSL)) as well as those on which Schedule 8 of the track access agreements is based. More generally, monitoring of the metric must support rather than undermine the delivery of other targets for which the routes will be accountable in CP6.

iv. **Customer focus** - the metric must reflect the experience of the passenger (and the freight customer). It must therefore change in line with the lateness that they experience when the services that they use are subject to performance failures on the route in question.

v. **Credibility** - the metric must be acceptable to, and capable of being understood by, relevant stakeholders, including route directors, passenger and freight train operators and customer representative organisations. To some extent, how well the metric meets this objective will depend on how far it meets the previous four. However, the need for stakeholders to accept and understand the metric also implies a degree of simplicity, recognising the lessons learned from experience with other regulatory metrics that have been subject to challenge in the past. In particular, we note that the possession disruption index failed to gain credibility due the complexity of the associated calculation and the difficulty of interpreting the results, and it is important that the credibility of route level performance monitoring is not similarly undermined.

Trade-offs and priorities

3.5 We note that there is some tension between these objectives. For example, the need to make fair comparisons between routes in the interests of ensuring clear accountability (objective i) tends to suggest a metric based on Minutes Delay (MD), which are captured at a large number of Recording Points located across the network and attributed to cause codes through the investigation of underlying incidents. At the same time, full alignment with the lateness experienced by passengers (objective iv) might imply a metric based on measures underpinning Schedule 8 of the track access agreements such as Performance Minutes (PM) or Average Minutes Lateness (AML), which are designed to be used in combination with estimates of passenger behaviour in response to lateness to generate revenue impacts. Lateness is captured at a series of Monitoring Points located at key destination stations, and weighted according to an estimate of the number of passengers alighting at the station in question.

3.6 Similarly, the need to ensure comparability (objective i) may conflict to some extent with ease of calculation (objective ii). This is because taking account of factors affecting performance that a route cannot reasonably be expected to manage is likely to involve a more complex calculation, for example modification of simple MD data to reflect the actual contribution of a route to managed performance before undertaking the final calculation. This issue is discussed further in Chapter 4, but here we note that such complexity, while it may be essential from the perspective of the routes to enable proper comparability, may also be less attractive to train operators and other stakeholders seeking to understand and interpret the metric.

3.7 Hence, in practice the chosen metric is likely to represent a compromise, meeting some objectives fully but others only partially. For the purposes of appraising the metrics specified and reviewed later in this report, we propose prioritising the objectives according to the order in which they appear in paragraph 3.4. We provide a rationale for this prioritisation in the table below.

Table 3.1: Prioritisation of objectives

Objective	Rationale
Comparability	This is a fundamental requirement that the metric must meet if it is to enable effective monitoring of route level performance and provide clear incentives to which the routes can respond. If it fails to meet this objective, it would arguably be better to rely on scorecard metrics rather than make comparisons across routes on the basis of potentially misleading measurement.
Ease of calculation	The ORR wishes to introduce the metric in CP6 (and possibly earlier on a trial basis), which precludes the development of new processes and systems for data collection and analysis. This objective is therefore also a basic requirement and failure to meet it would rule out the introduction of a metric at least in the short to medium term.
Consistency	The introduction of route scorecards is an important benefit of devolution to the route level, while the incentives within Schedule 8 of the track access agreements are well established and are unlikely to change (materially) at the start of CP6. The new route level performance metric must therefore work in tandem with these measures if the regulatory framework is to encourage appropriate behaviours on the part of the routes. At the same time, given the variety of metrics included in the scorecards, there must be some recognition that performance outcomes measured in different ways will sometimes diverge.
Customer focus	The metric must align with the customer experience, but this does not necessarily mean that it should capture lateness incurred by them directly. In practice, a route will have only limited influence on the distribution of passengers across services, and a metric that captures this arguably does not reflect the route’s contribution to performance. Moreover, as we will show in Chapter 4, lateness and delay tend to be highly correlated. In our view, it is appropriate to rely on this correlation to meet this objective rather than compromising on the first three.
Credibility	It is unlikely that all stakeholders will consider that a given metric represents the best option, since the criteria that they apply in selecting their preferred option will vary. We suggest that it is most important that the routes understand and support the specification of the metric and that the ORR is confident that it captures the performance for which the routes can be held accountable. Moreover, the routes, the ORR and train operators already have the technical understanding to monitor and interpret different aspects of performance. Hence, in our view, the introduction of a simple metric that has widespread support among the broader stakeholder community but which is not accepted as a fair basis for comparison by the routes would seriously undermine its value. Moreover, regardless of the specification chosen, the introduction of the metric should anyway be supported by the dissemination of information on how it is calculated and interpreted. This would help to build credibility without the need to compromise unduly on other objectives.

Source: Steer Davies Gleave

4 Issues in metric design

Route contribution to performance

4.1 Given the objective of comparability discussed in the previous chapter, it is important to consider the role that the routes have in delivering performance since this must inform any understanding of the performance for which they are accountable. From a performance perspective, this role can be defined as follows:

- **Asset stewardship** – this involves developing and implementing an asset management strategy with a view to avoiding as far as possible, and otherwise minimising, asset failures that could disrupt the performance of train services on the relevant part of the national rail network[2].
- **Timetable planning** – the routes support Network Rail's system operator function in specifying the Train Planning Rules governing the planning of the scheduled services that run over their territory. The reliability of a given timetable will generally depend on how well these rules are specified.
- **Control according to the timetable** – the route is also responsible for making day-to-day operational decisions that enable train services to run in accordance with a published timetable and hence without unplanned disruption.
- **Management of incidents occurring on the route** – when incidents do occur, whether they are the result of asset failures or operational decisions (or a combination of the two), they must be managed in order to minimise the impact on train services and their customers. In practice, this is an important element of the role, especially on parts of the network that are used intensively and there is a need to coordinate the incident responses of a wide range of parties.
- **Mitigation of delay imported from other routes** – some of the delay experienced by trains operating on a given route may be the result of incidents arising on other routes, the impact of which is exported across one or more route boundaries. Such delay must also be mitigated, which typically requires effective communication with the route that has caused it as well as train operators affected by it.

4.2 An indication of how well a route is performing in each element of the role is provided by the different types of delay captured by the existing recording and monitoring systems. More specifically, the route's performance in terms of asset stewardship and control, and to some extent in terms of timetable planning and incident management, is captured by primary delay

[2] Notwithstanding the focus of our work on performance, we note that asset stewardship is clearly essential for the safety of the railway.

due to incidents occurring within its territory[3]. Its timetable planning and incident management capability can also be measured by reference to the secondary delay arising from such incidents, while its ability to mitigate imported delay is captured by secondary delay incurred on the route in question but attributable to incidents arising elsewhere[4].

4.3 This demonstrates the difficulty of defining a metric that accurately captures the performance for which a route is genuinely accountable while keeping the calculation simple. In particular, it shows that basing the calculation on the level of delay attributed to a route, or alternatively that incurred on the route, will not result in an accurate measure of a route's contribution to performance. Under the first approach, the calculation fails to take account of the route's ability to mitigate imported delay while overstating its ability to influence delay incurred on the opposite side of a route boundary. Under the second, the calculation excludes some delay caused by the route although it does include imported delay that it is in a position to influence.

4.4 We consider how to address the issue of imported delay further in the specification of potential metrics described in the following chapter. Here, we note that there may be a case for introducing different metrics representing different elements of a route's role rather than seeking to combine all aspects of its contribution to performance into a single measure.

Capturing the passenger experience

Lateness and delay

4.5 The lateness of train service depends in part on how successfully the route undertakes the role defined in paragraph 4.1. Moreover, lateness is arguably a better measure of success than delay, since it reflects the impact of sub-threshold minutes that are not captured by the delay attribution process as well as correlating more closely with the experience of passengers and users of freight services. However, as discussed further in the following chapter, there is currently no basis for allocating lateness between routes, since the Monitoring Points used to capture lateness do not align with route boundaries. While Network Rail has devised a method of allocation based on delay attribution, this is not supported by performance reporting and monitoring systems as they are currently configured, and it is anyway not clear whether it would generate estimates of route level performance that were genuinely comparable across routes.

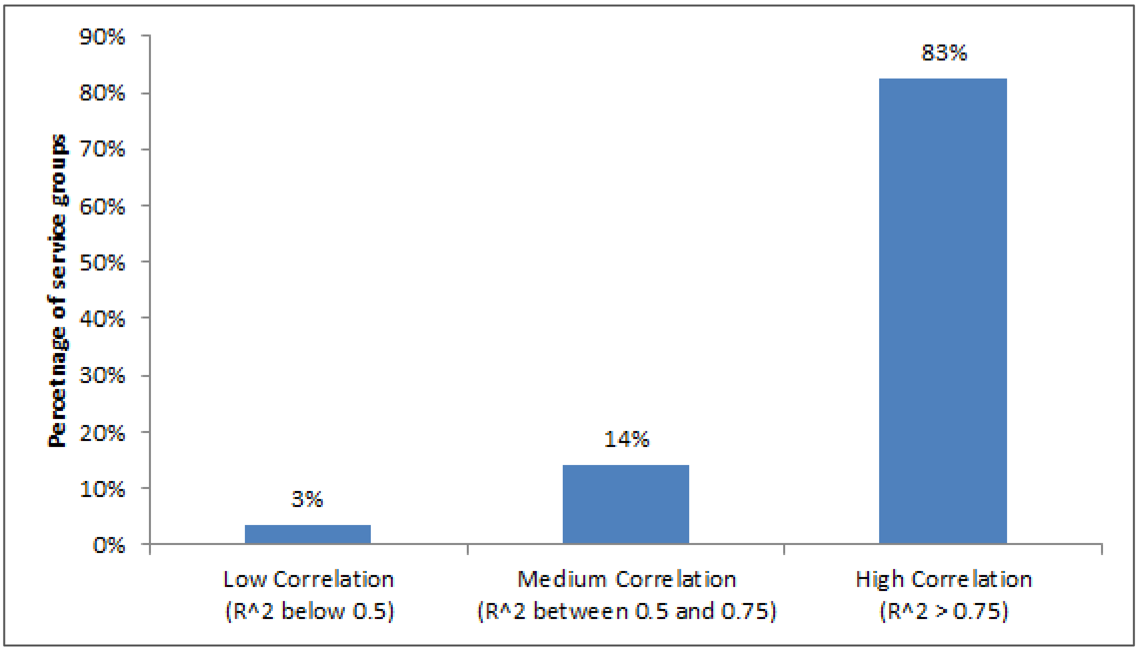
4.6 In any event, MD are generally highly correlated with lateness, as shown in Figure 4.1. This means that any metric based on MD will tend to reflect the passenger experience, even if it does not capture lateness precisely. As already discussed in the course prioritising the objectives in Chapter 3, there is a case for relying on this correlation rather than seeking to base route level measurement on a metric such as lateness that does not align well with route geography.

---

[3] Primary delay is defined as the delay to a train arising directly from an incident.

[4] Secondary delay, otherwise known as reactionary delay, is defined as delay to a train arising indirectly from an incident previously affecting the same train or other trains.

Figure 4.1: Correlation between Minutes Delay and lateness



Source: Steer Davies Gleave analysis based on Network Rail PEARS data

Weighting by passenger numbers

4.7 An alternative to using lateness as a measure of performance would be to weight MD by passenger numbers, for example by deriving weights from the volume of passengers using differ Service Groups operating on each route. Such a weighting exercise could provide useful information if the aim is to prioritise the allocation of performance improvement resources with a view to reducing passenger disruption overall. More specifically, it could inform decisions about the trade-off between capacity and performance, the appropriate timing of interventions and how best to regulate services in real time. This approach is considered further in the context of comparability and normalisation discussed in the following section.

Ensuring comparability

4.8 As already noted, routes differ markedly in their scale and technical characteristics as well as the types of train service that operate on them. This means that in order to compare performance by route, however defined, it is necessary to normalise the measure using an appropriate scaling or other factor. In principle, a number of normalisation factors are available, as follows:

- Measures of the scale of the infrastructure: these are measures of the physical size of the infrastructure, such as route km or track km, which give a broad indication of both the scale of the route’s responsibilities and the potential for service disruption (a larger network will require more resources to operate and manage and will tend to support more services generating more delays). There is therefore a case for measuring route performance in terms of delay per route or track km.
- Measures of the level of service activity: alternatively, it is possible to normalise a route performance measure using a factor that directly captures the level of service activity, such as train km or vehicle km. In broad terms, the greater the number of train km operated, the greater the potential for delay to services and the greater the total level of delay is likely to be. Measuring route performance in terms of delay per train km

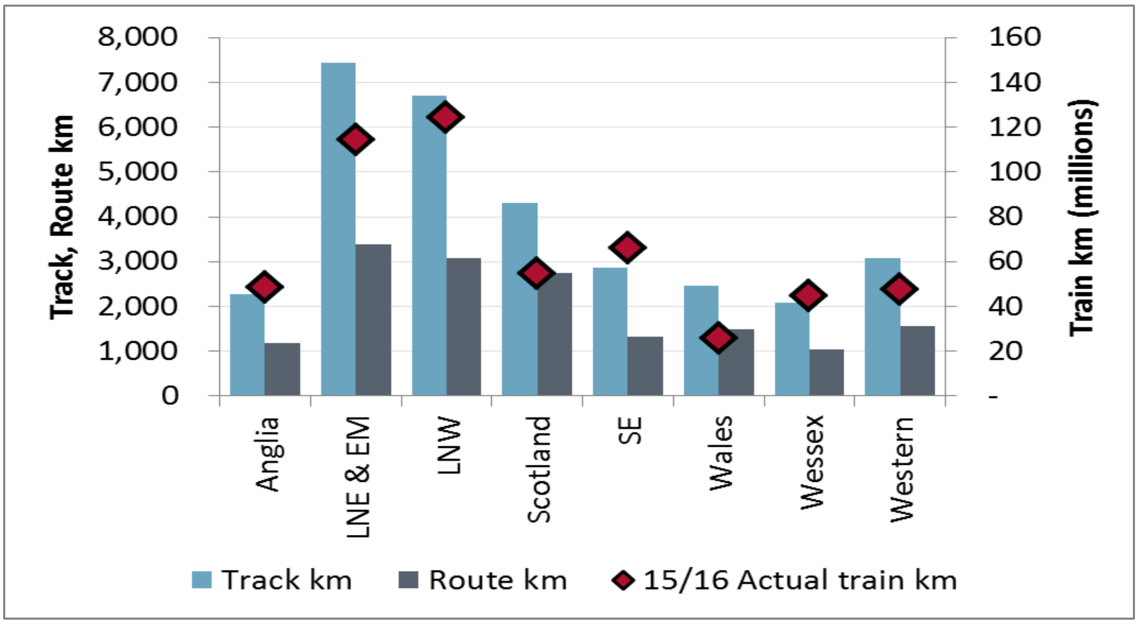
therefore enables a more meaningful comparison than delay per track km, since it reflects a key driver of performance over which routes generally have little control.

- Measures of the level of passenger traffic: route performance can also be expressed in terms of delay per passenger journey or per passenger km. Such a measure directly captures the passenger experience, particularly when combined with a passenger-weighted measure of delay. We understand that there is currently no allocation of passenger km to routes, but that an allocation of passenger journeys, albeit imprecise, is available.
- Measures of the level of service intensity: it is also possible to combine normalisation factors to reflect different characteristics of a route affecting the level of performance. In particular, dividing train km by track km gives a measure of service intensity (reflecting the volume of traffic in relation to the scale of the infrastructure). This may be a more important driver of delay than the level of service activity, since the more intensively the infrastructure is used, the more likely a given incident will lead to propagation of delay across different services. Again, the level of service intensity cannot be materially influenced by the route, and there is therefore a case for controlling for it in the design of the metric.

4.9 We have also considered other factors that could be taken into account in normalising a given metric, such as the degree and type of electrification, the extent of double or four-line tracking and the share of freight traffic in the total for the route. However, we have concluded that it is not appropriate to control for all the factors that could contribute to the level of performance observed, since this would unduly complicate the metric and make it difficult to interpret.

4.10 The figure below shows the relationship between three potential normalisation factors, namely track km, route km and train km (with train km derived from actual operations in 2015/16). It indicates that all three measures are broadly correlated, with routes with a higher proportion of longer distance service – London North Eastern and East Midlands (LNE & EM) and London North Western (LNW) - accounting for proportionally more train km notwithstanding the greater service intensity in the South East (SE).

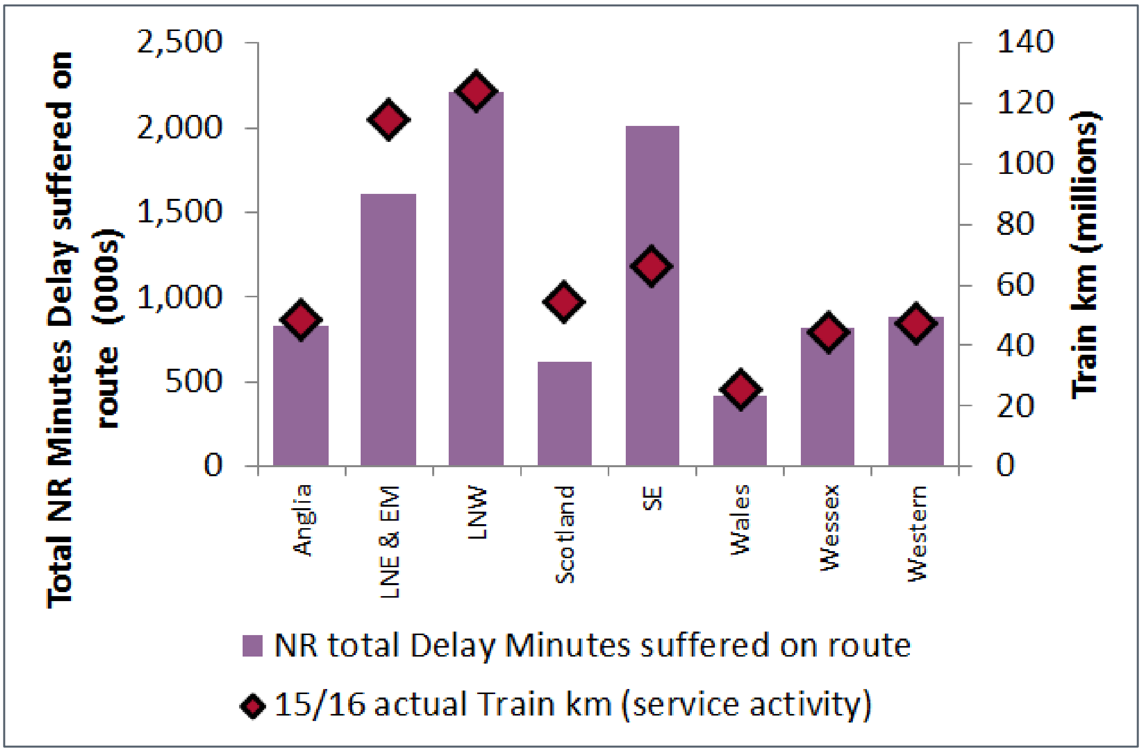
Figure 4.2: Potential normalisation factors



Source: Steer Davies Gleave analysis based on Network Rail data

4.11 However, while the correlation between these factors suggests that any one might provide a means of scaling a given measure of performance, we consider that train km is the most appropriate. This is because performance, when measured using a metric such as total MD, is driven to some degree by the number of trains operating on a given part of the network, as illustrated in the figure below. Overall, the figure suggests a strong relationship between train service activity and delay, although the performance of SE is substantially worse than might be expected from observation of the associated level of traffic alone (and the performance of LNE & EM and Scotland somewhat better).

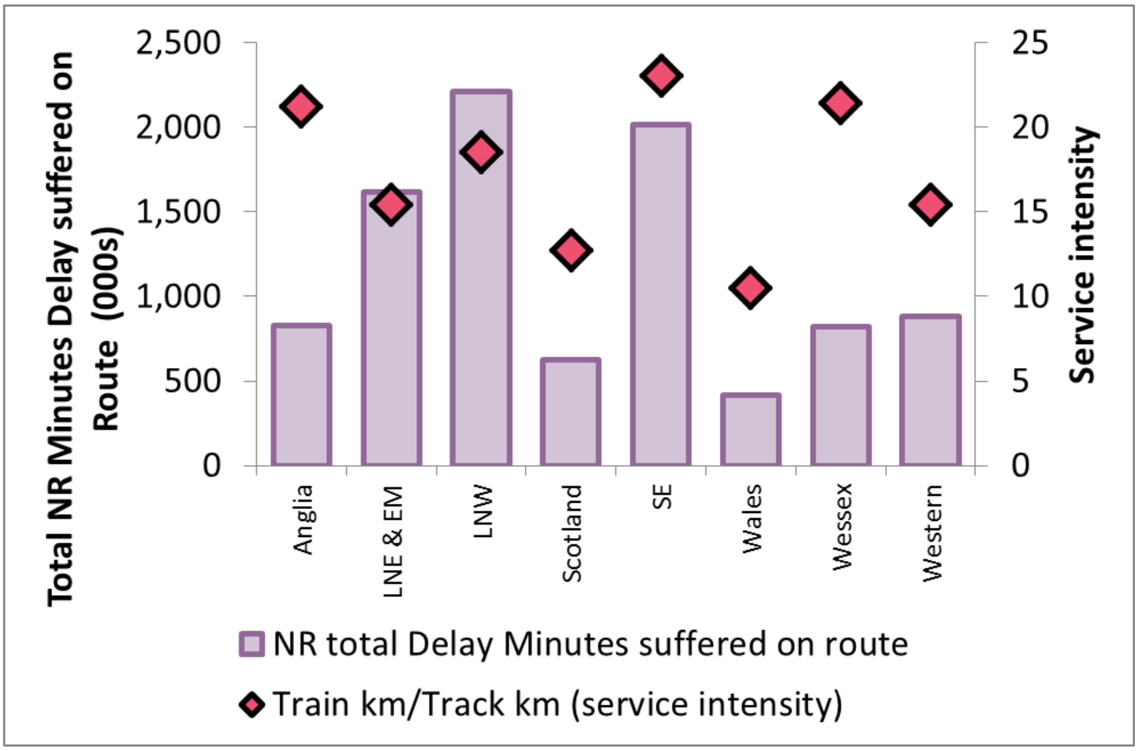
Figure 4.3: Relationship between train service activity and delay



Source: Steer Davies Gleave analysis based on Network Rail data

4.12 Figure 4.4 shows the relationship between service intensity and delay, and arguably provides an explanation of the relatively poor performance of SE when measured in terms of total delay. SE is the most intensively used part of the national rail network, with many high frequency services sharing the same lines, particularly during peak periods. Less intensively used routes, such as Scotland and Wales, experience much lower levels of delay. At the same time, Anglia and Wessex, both of which support frequent commuter services into London in common with SE, appear to perform relatively well on the basis of this comparison.

Figure 4.4: Relationship between train service intensity and delay 2015/16



Source: Steer Davies Gleave analysis based on Network Rail data

4.13 In our review of possible metrics set out in the following chapter, we have investigated the application of different normalisation factors depending on the performance measure in question. Where appropriate, we have calculated the metric in two different ways, applying both service level (train km) and service intensity (train km per track km) in order to compare normalised results. In principle, we consider service intensity to be the more appropriate basis for normalisation as it takes account of more information about the characteristics of a route. At the same time, we note that a metric such as MD per unit of service intensity is arguably less easy to interpret than MD per train km. These issues are considered further as part of the appraisal of metrics reported in Chapter 7.

5 Specification of the metrics

Approach

5.1 We have specified six metrics for further investigation, taking account of the objectives described in Chapter 3 and the design issues discussed in Chapter 4. These include three already identified by the ORR in the course of its previous work and in discussion with Network Rail. The metrics selected are as follows:

- Minutes Delay (M1), based on the total number of Network Rail-caused MD attributed to a route, with attribution determined according to the location of the underlying incident;
- Passenger weighted Minutes Delay (M2), whereby the calculation for M1 is modified in order to weight MD by passenger journeys;
- Performance Minutes (M3), based on those already monitored for the purposes of Schedule 8 of the track access agreements, modified to capture a route's contribution to performance;
- Primary Minutes Delay (M4), which focuses on Network Rail-caused primary delay (defined according to the Delay Attribution Guide) attributed to routes in the same way as for M1, and a variant ($M4^+$), which includes primary and an element of secondary delay;
- Secondary Minutes Delay (M5), based on the total level of secondary delay incurred on the route, regardless of the party causing it; and
- Deemed Minutes Lateness (M6), a measure of the reliability of the service that also relies on concepts and parameter values used in Schedule 8 of the track access agreements.

5.2 In the following paragraphs, we set out the specification of each metric in more detail and summarise the key issues that it raises. Note that the metrics defined here are not necessarily mutually exclusive, and in principle could form a suite of KPIs to inform discussions between the routes and their customers as well as regulatory decisions. However, as noted in paragraph 3.3, the appraisal reported in the final chapter has been undertaken on the assumption that the metric will be used as a basis for setting formal regulatory targets.

5.3 In all cases, the metric can be calculated for any period of time recognised by standard railway accounting (e.g. a four-week period or a year). The appropriate periodicity for the monitoring framework is not considered further here, although we anticipate that cross-route comparisons of performance levels would be made periodically, while comparisons of rates of change could be on the basis of average movements over a year or more. Network Rail have confirmed that historical data within their performance database has been remapped to current route boundaries, and that it would be possible to obtain the data used to calculate the metrics described here from 2009 onwards.

5.4 In the case of metrics M1, M4, $M4^+$ and M5, the specification includes both passenger and freight delay (although they could be specified as passenger-only metrics if required). M2, M3 and M6 are passenger-focused metrics and therefore exclude freight-related performance. Note that in the formal specification of the metrics provided below, we refer to Service Groups for the purposes of simplicity but the terminology can be broadened to cover both freight and passenger services.

Metric definitions

M1: Minutes Delay by Route (passenger and freight delay combined)

5.5 This metric can be defined simply as the total number of Network Rail-caused MD attributed to a route, with attribution based on the location of the underlying incident. The formulation of the metric is as follows:

$$M1 = \frac{\sum_{i,s} MD_{i,s}}{NF},$$

Where:

- $MD_{i,s}$ is the Minutes Delay incurred by Service Group 's' in a given period as a result of incident 'i' caused by a given route;
- NF is a normalisation factor given by $\sum_s TK_s$ (train service activity) or $\frac{\sum_s TK_s}{RK}$ (train service intensity), with:
  - TK_s equal to the total number of train km operated by Service Group 's' on the route over the relevant period; and
  - RK equal to the total number of track km for the route.

5.6 Under this simple formulation, any secondary delay exported across a route boundary would be attributed to the route on which the original incident arose. Hence, while it captures performance for which the route can be considered accountable, it fails to take account of a route's inability to mitigate delay once a train has passed into the territory of an adjacent route. This means that normalisation based on train km operated on the route may be misleading, since a portion of the delay included in the metric is accumulated over train km not captured by the normalisation factor. Indeed, some trains affected by a given incident may operate entirely on adjacent routes, in which case they will be entirely excluded from the train km used for the purposes of normalisation. The significance of this weakness in the calculation will depend on the extent to which delay is exported in practice, an issue considered in the following chapter.

M2: Passenger Weighted Minutes Delay by Route (freight performance excluded)

5.7 In this formulation, the Minutes Delay incurred by each Service Group affected by incidents on the route in question are weighted by the journeys undertaken on the Service Group. The formulation is given below:

$$M2 = \frac{\sum_{i,s}(MD_{i,s} \times J_s)}{\sum_s J_s},$$

Where:

- J_s is the number of journeys on Service Group 's' in a given time period;
- $MD_{i,s}$ has the meaning given to it in the definition of M1 above.

5.8 This metric is subject to the same weakness as M1 in that it takes no account of a route's inability to mitigate delay exported beyond its boundary, and the same limitations in respect of normalisation therefore apply. In addition, since the value of the metric would be affected by the number of journeys on individual Service Groups, observed changes over time might be the result of factors outside the route's control (although the weightings could be expected to remain constant over a five-year CP).

M3: Performance Minutes by Route (freight performance excluded)

5.9 Network Rail has proposed a metric that is based on measurement of lateness rather than delay and is aligned with the metrics used in Schedule 8 of the track access agreement. This would involve allocating the lateness incurred by individual Service Groups to routes using the attribution of MD, in much the same way that lateness is currently allocated between Network Rail and train operators according to the party responsible for the underlying incident. Based on a worked example prepared by Network Rail, we have identified the following formulation, although this has not been discussed and confirmed with the infrastructure manager:

$$M3 = \frac{\sum_{m,s}\left[\left(ML_{m,s}\times\sum_i\{MD_{i,s}\div MD_s\}\right)+\left(\sum_i RC_{i,m,s}\times CM_s\right)\right]\times MPW_{m,s}}{\sum_{m,s} SP_{m,s}},$$

Where:

- $ML_{m,s}$ is the Minutes Lateness for all trains in Service Group 's' at Monitoring Point 'm' over a given time period;
- $MD_{i,s}$ has the meaning given to it in the definition of M1;
- MD_s is all Minutes Delay incurred by Service Group 's' over the relevant time period;
- $RC_{i,m,s}$ is the number of Cancelled Stops at Monitoring Point m incurred by Service Group 's' due to incident i caused by a given route;
- CM_s is the Cancellation Minutes for Service Group 's' specified in the relevant track access agreement;
- $MPW_{m,s}$ is the weighting for Monitoring Point m and Service Group 's'; and
- $SP_{m,s}$ is the number of scheduled stops at Monitoring Point m made by Service Group 's'.

5.10 This formulation draws heavily on the calculations supporting the determination of performance payments in Schedule 8 of the track access agreement. The key difference is that it apportions lateness to an individual route according to the portion of MD and Cancelled Stops for which the route is responsible. However, in line with the approach taken for the purposes of Schedule 8, there is no mechanism for matching individual lateness and delay records, and it is therefore unclear how far the metric reflects route accountability. Moreover, although it captures average lateness for all Service Groups affected by route performance failures, it does not provide for normalisation by reference to route characteristics, raising further doubts about whether it would support a fair comparison of route level performance.

5.11 We also understand that calculation of the metric on a regular basis would require significant modification to Network Rail's PEARS system. While a rewrite of PEARS is currently planned, the timing of this is unclear, and there is therefore some uncertainty over whether the necessary changes could be implemented in time to enable the introduction of the metric at the beginning of CP6.

M4: Primary Minutes Delay by Route (passenger and freight delay combined)

5.12 An option for addressing the shortcomings of M1 and M2 relating to normalisation is to calculate a metric based on primary delay only, as in the following formulation:

$$M4 = \frac{\sum_{i,s} PRIMD_{i,s}}{NF},$$

Where:

- $PRIMD_{i,s}$ is the primary MD incurred by Service Group ‘s’ as a result of incident ‘i’ caused by a given route; and
- NF has the meaning given to it in the definition of M1 above.

5.13 This would ensure that only MD for which a route was fully responsible were included in the calculation of its performance. Normalisation using train km operated or a measure of train service intensity within the relevant route boundaries would then be more meaningful since, by definition, all secondary delay exported across route boundaries is excluded from the calculation.

$M4^{+}$: Primary and Secondary Minutes Delay caused by and incurred on a given Route (passenger and freight delay combined)

5.14 However, M4 fails to capture all of the delay that a route might be expected to manage or mitigate. This argues for an alternative specification, whereby it is modified to include primary delay caused by the route and secondary delay that is both caused by and suffered on the route[5]. In this specification, the numerator therefore excludes all delay imported and exported across route boundaries, but nevertheless captures much of the secondary delay that the route must mitigate.

$$M4^{+} = \frac{\sum_{i,s}(PRIMD_{i,s} + SERMD_{i,s})}{NF},$$

Where:

- $SERMD_{i,s}$ is the secondary MD incurred by Service Group ‘s’ as a result of incident ‘i’ caused by a given route; and
- The other terms have the meaning given to them in the definition of M4 above.

5.15 The formulation of M4 could also be changed to allow for weighting by passenger journeys, as in M2. This would make it more representative of the passenger experience, but would also add to the complexity of the calculation.

M5: Secondary Minutes Delay by Route (passenger and freight delay combined)

5.16 As a supplementary measure to M4, it would also be possible to calculate a metric capturing a route’s performance in terms of managing all secondary delay within its boundaries, regardless of the party causing it and where it was caused. This would require measuring a route’s MD according to where it occurred, as in the following formulation:

---

[5] For the avoidance of doubt, for a given minute of delay to be included in the numerator for $M4^{+}$, it must be both caused by an incident occurring on the route and suffered by a train travelling on the route.

$$M5 = \frac{\sum_{r,s} SERMD_{r,s}}{NF},$$

Where:

- $SERMD_{r,s}$ is the secondary MD incurred by Service Group 's' at Recording Point r on the route in question; and
- *NF* has the meaning given to it in the definition of M1 above.

5.17 Again, normalisation through the application of a measure of train service activity or intensity would arguably be more meaningful than in the case of M1, since the delay captured by M5 occurs within the same boundaries used to determine the train km and track km data on which the normalisation factor is based. However, by definition the metric excludes a significant proportion of the MD for which the route is accountable, while including a substantial proportion that it has not caused.

M6: Deemed Minutes Lateness by Route (freight performance excluded)

5.18 The metrics defined above that are based on MD focus on train service punctuality and exclude any measure of reliability from the calculation[6]. Hence, were any of these metrics to be selected, it should arguably be supplemented with a measure of reliability. Such a measure could be based on the Cancellation Minutes element of M3, as follows:

$$M6 = \frac{\sum_{i,m,s} RC_{i,m,s} \times CM_s \times MPW_{m,s}}{\sum_{m,s} SP_{m,s}},$$

Where each term has the meaning given to it in the definition of M3 above.

5.19 The metric is subject to the same strengths and weaknesses as M3.

[6] The TRUST 'Delay' dataset is a repository of both delay minutes and cancellations. Cancellations are converted into deemed minutes using the Schedule 8 parameters, which incorporate the effect of the service frequency interval on passenger lateness. For each incident, the root cause and delay/deemed minutes is recorded as well as the location where the incident occurred and where each service is affected. The use of this dataset could enable any delay minute metric to be expanded to also include cancelled services. However, this would require further investigation of TRUST, which we have not sought to undertake as part of this study.

6 Review of the metrics

Route comparisons

6.1 We have calculated each of the six metrics defined in the previous chapter for all routes using data for 2015/16 provided by Network Rail. The results are presented below. Note that in all cases, the calculation relies on standard industry data available for a number of years, and that historic trends for each metric could therefore be generated. However, as already noted, in some cases industry systems do not yet support the necessary calculation, notwithstanding that the required data exists in principle[7].

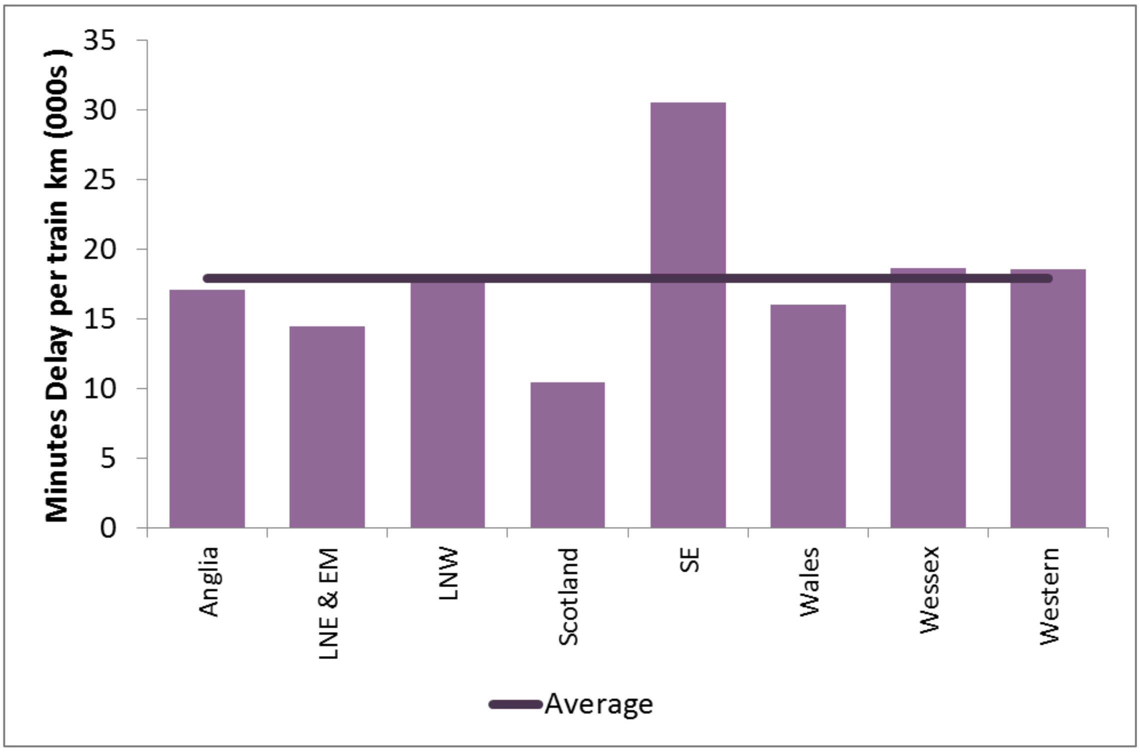
6.2 In the case of metrics M1, M4 and M5, we have applied both of the normalisation factors described in Chapter 4 (train service activity and train service intensity) in order to illustrate the impact of different approaches to improving comparability across routes. In each case, the first figure shown illustrates route level performance normalised according to train service activity (train km) and the second according to train service intensity (train km per track km). As already noted, metrics M3 and M6 are derived from formulae within Schedule 8 of the track access agreement and, while they provide a measure of average performance, are not subject to further normalisation. M2 is a passenger-weighted metric and is therefore normalised using a measure of passenger journeys. Hence, the results for each of these three metrics are presented in a single figure.

M1: Minutes Delay by Route

6.3 As shown in Figure 6.1, the performance of most routes when measured according to delay caused per train km is broadly similar, and only SE's is substantially worse than the average (with performance in Scotland significantly better than average). As already noted, one objection to this comparison is that it does not take account of the greater service intensity on SE, which tends to increase the propagation of delay between services following any given incident. This can be seen from Figure 6.2, which shows the same MD data normalised by the ratio of train km to track km. Measured on this basis, SE's performance is closer to the average for all routes, while the longer distance routes stand out in terms of above-average delay. In our view, the comparison in Figure 6.2 is more meaningful, since it controls to some degree for service intensity, a characteristic over which the routes arguably have only limited influence.

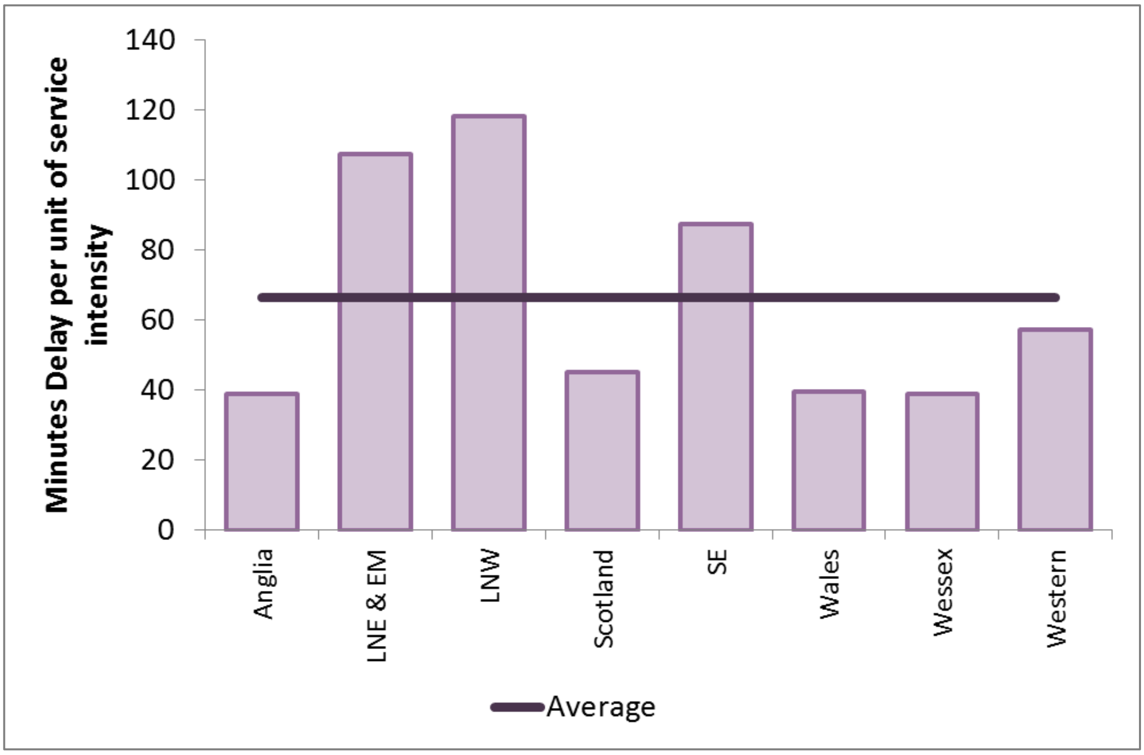
[7] We understand that is the changes to systems required in order to calculate M3 and M6 were made, it would be possible to generate historic data series.

Figure 6.1: Network Rail-caused Minutes Delay per train km 2015/16



Source: Steer Davies Gleave analysis based on Network Rail data

Figure 6.2: Network Rail-cause Minutes Delay per unit of service intensity 2015/16

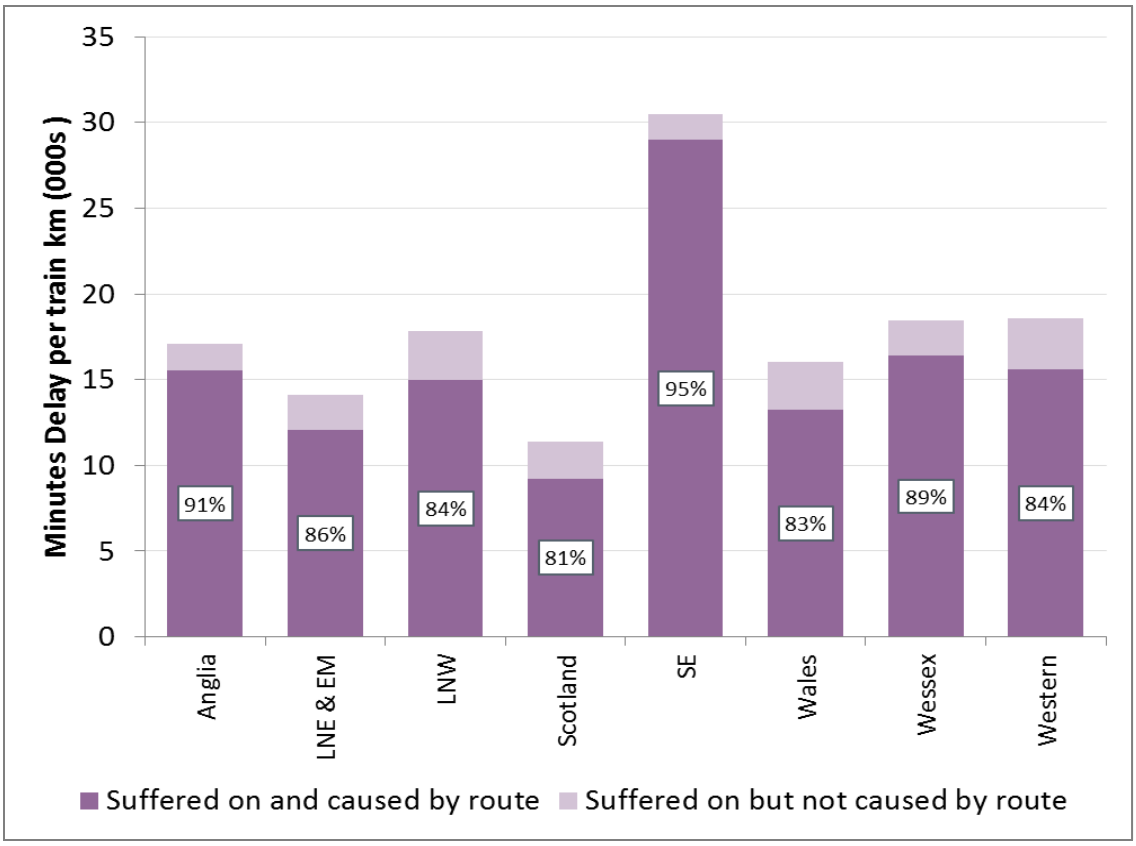


Source: Steer Davies Gleave analysis based on Network Rail data

6.4 However, both of the metrics illustrated above are subject to the limitation discussed in the previous chapter concerning the export of delay. As indicated there, the allocation of all MD to the route on which the underlying incident occurs takes no account of a route's inability to manage secondary delay incurred by a train that has passed into the territory of an adjacent route. In practice, the impact of this limitation on comparability depends on the extent to which delay is imported/exported across route boundaries, an issue that we have investigated using data from Network Rail.

6.5 In Figure 6.3, we show the proportion of delay suffered on a route but not caused by it, which can be described as imported delay. As indicated, such delay accounts for more than 10% of the total on six routes and more than 15% on four. In our view, these are sufficiently high proportions to distort a measure of performance that is based on an allocation of delay according to where it is caused but normalised by reference to service activity or intensity within defined route boundaries. Moreover, we note that they could become higher if route boundaries or service patterns change (although we recognise that they might also become lower). We therefore conclude that metric M1, as defined, is likely to be open to challenge on the grounds that it does not support a fair comparison of performance between routes.

Figure 6.3: Extent of Network Rail-caused Minutes Delay generated and imported by route 2015/16

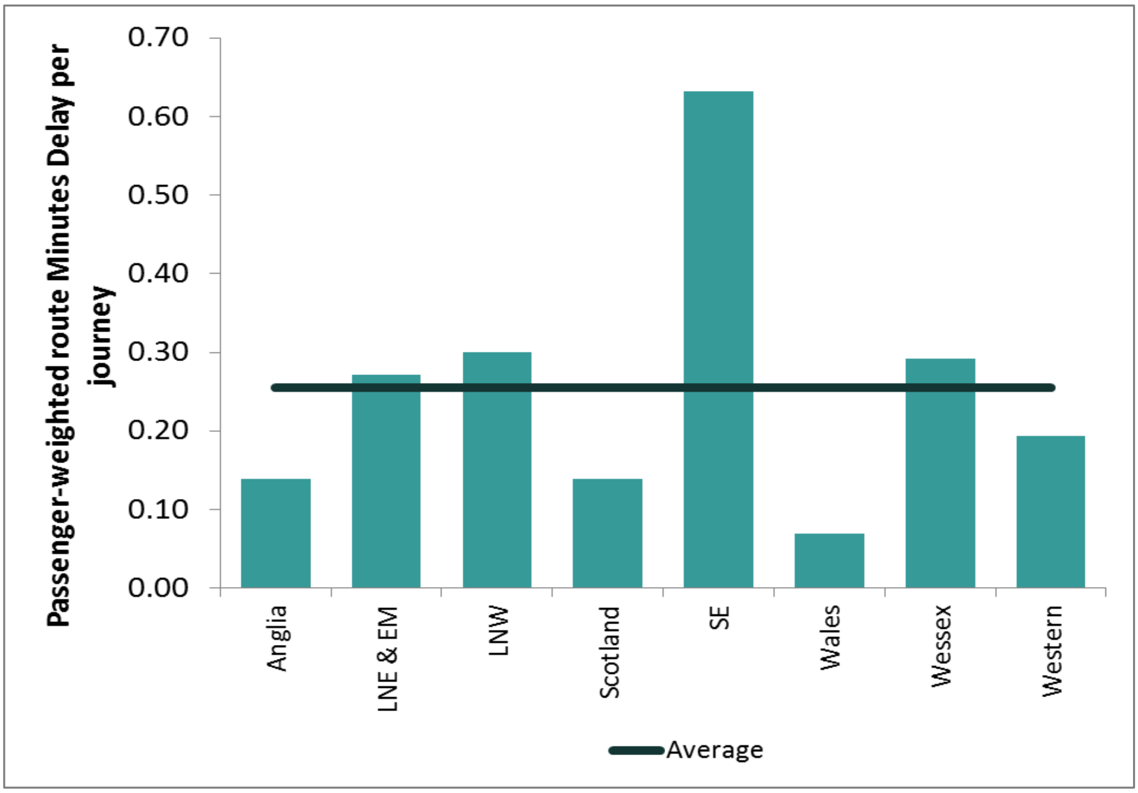


Source: Steer Davies Gleave analysis based on Network Rail data

M2: Passenger Weighted Minutes Delay by Route

6.6 Figure 6.4 illustrates the effect of weighting delay by passenger journeys for the relevant Service Groups by comparing metric M2 with M1. As indicated, the introduction of weighting does not change the results of the comparison substantially, although performance in Scotland and Wales appears better under M2 as the number of passengers using the services on these routes is less on average. The apparently relatively poor performance of SE continues to stand out however.

Figure 6.4: Network Rail-caused passenger-weighted Minutes Delay per journey 2015/16



Source: Steer Davies Gleave analysis based on Network Rail data

6.7 In our view, a cross-route comparison of performance on the basis of M2 is open to challenge since the value depends on the distribution of passengers across delayed services, an outcome driven by the service specification rather than any action that can be taken by the route. At the same time, the metric would provide useful information to a route seeking to focus resources in a way that delivers the maximum benefit for passengers. Other things being equal, a given reduction in train delay will result in a greater reduction in the value of the metric if it is achieved on a more heavily loaded service, encouraging route performance managers to focus their efforts accordingly. This suggests that the metric should be seen as a useful management tool rather than a basis for comparison within a formal regulatory framework.

6.8 Moreover, as noted in Chapter 4, there is no generally recognised basis for calculating passenger weighted delay. The results presented above are based on a relatively crude methodology, which relies on the allocation of total passenger journeys by Service Group to routes according to which routes are used by each Service Group. This may substantially distort the comparison, not least because it is based on the assumption that all passengers using a Service Group in a given year experience delay on any route on which the Service Group operates.

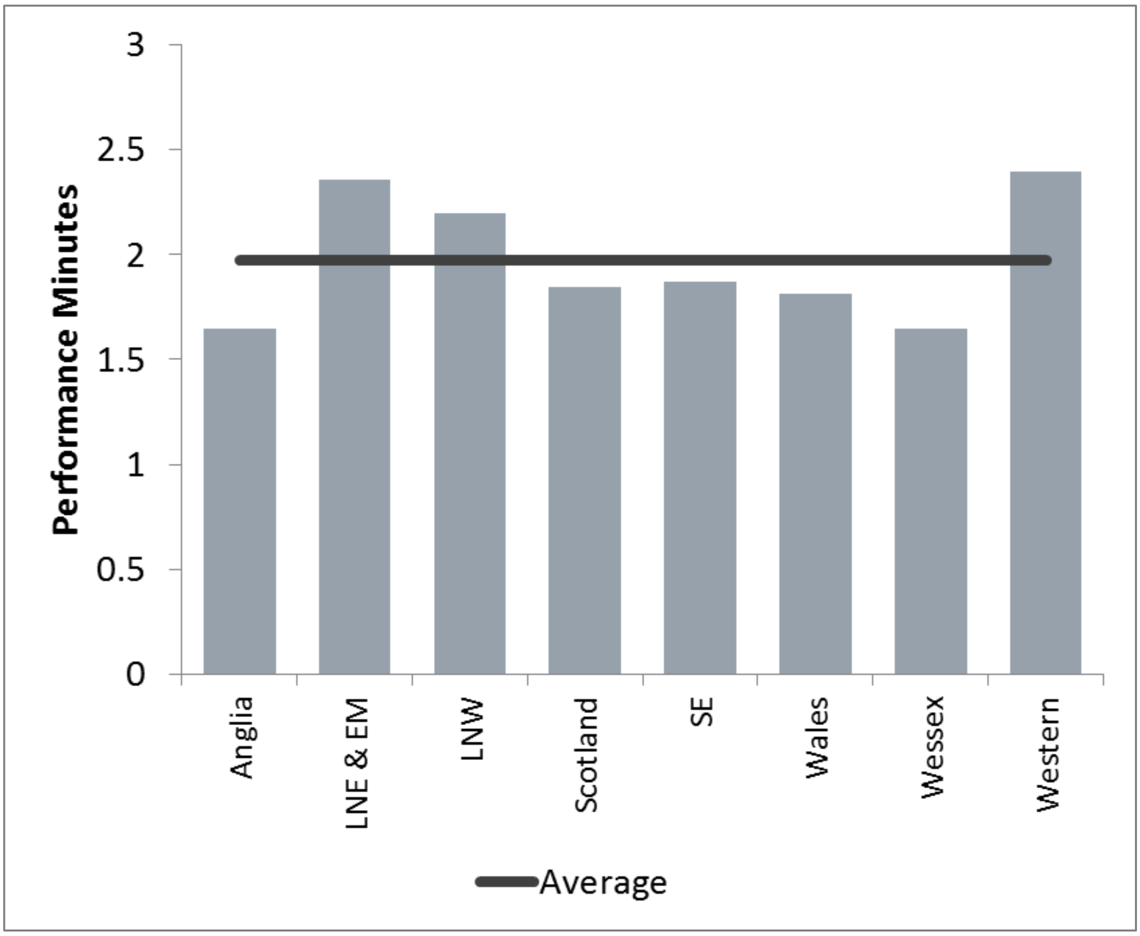
6.9 We therefore conclude that M2 does not provide an appropriate basis for comparing route level performance. It is possible that changes to data recording may enable a more robust passenger-weighted metric to be calculated in the future. Nevertheless, the use of such a metric as a basis for holding routes to account through a comparison of their performance levels is likely to be subject to challenge on the grounds that passenger journey numbers are largely determined by decisions taken by other parties.

M3: Performance Minutes by Route

6.10 As discussed in Chapter 5, it is not yet possible to calculate Performance Minutes by route in the way in which Network Rail has proposed and according to the specification provided in paragraph 5.9. We have nevertheless approximated the calculation by apportioning Performance Minutes by Service Group according to the proportion of MD incurred on different routes, and the results are shown in Figure 6.5. This suggests that the longer distance routes are the worst performing, possibly a reflection of the greater potential for delay to accumulate over a longer journey and hence the greater the level of lateness recorded for longer distance services at the point of destination. If correct, this explanation arguably provides more information about the performance characteristics of the services that run on each route rather than the contribution to performance of the routes themselves.

6.11 In any event, the interpretation of the metric is difficult because of the complexity of the calculation represented by the equation in paragraph 5.9, and further investigation would be needed to draw meaningful conclusions from this comparison. Given that it is not yet possible to make the calculation accurately, and in view of Network Rail's own observation that the lack of normalisation makes it difficult to use the metric as a basis for comparison, we conclude that it is not an appropriate metric for monitoring route level performance during CP6.

Figure 6.5: Network Rail Performance Minutes 2015/16

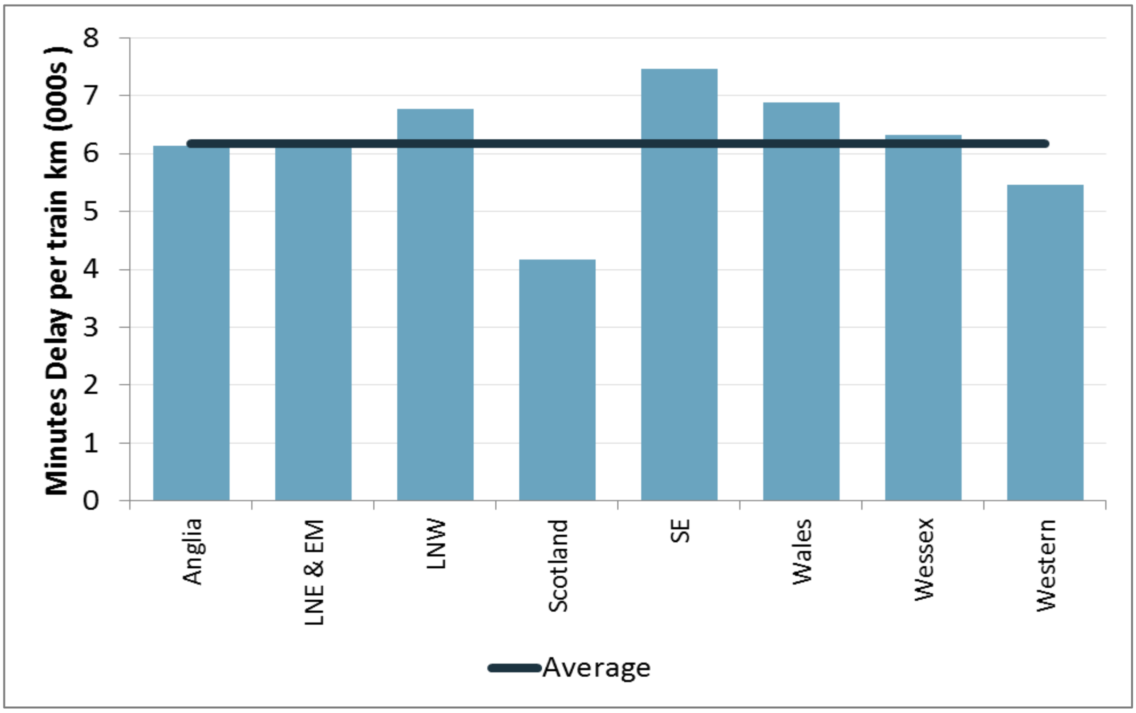


Source: Steer Davies Gleave analysis based on Network Rail data

M4: Primary Minutes Delay by Route

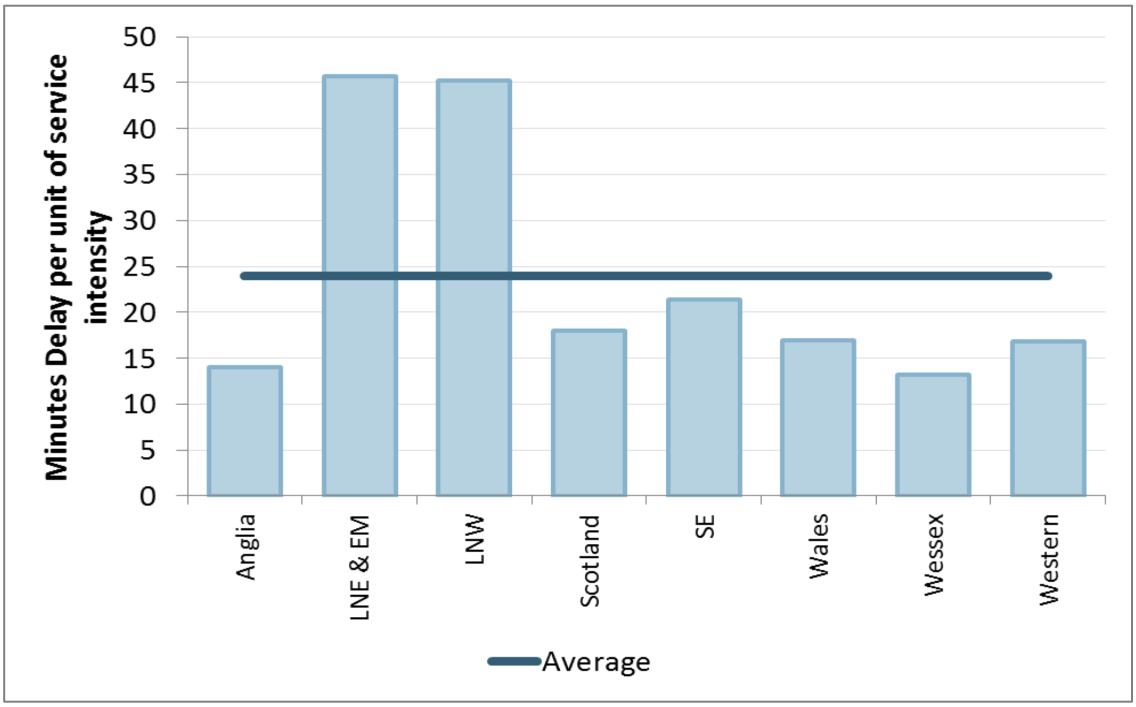
6.12 Figure 6.6 and Figure 6.7 provide a comparison of primary MD across the routes, normalised using train service activity and train service intensity respectively.

Figure 6.6: Network Rail-caused primary Minutes Delay per train km 2015/16



Source: Steer Davies Gleave analysis based on Network Rail data

Figure 6.7: Network Rail-caused primary Minutes Delay per unit of service intensity 2015/16



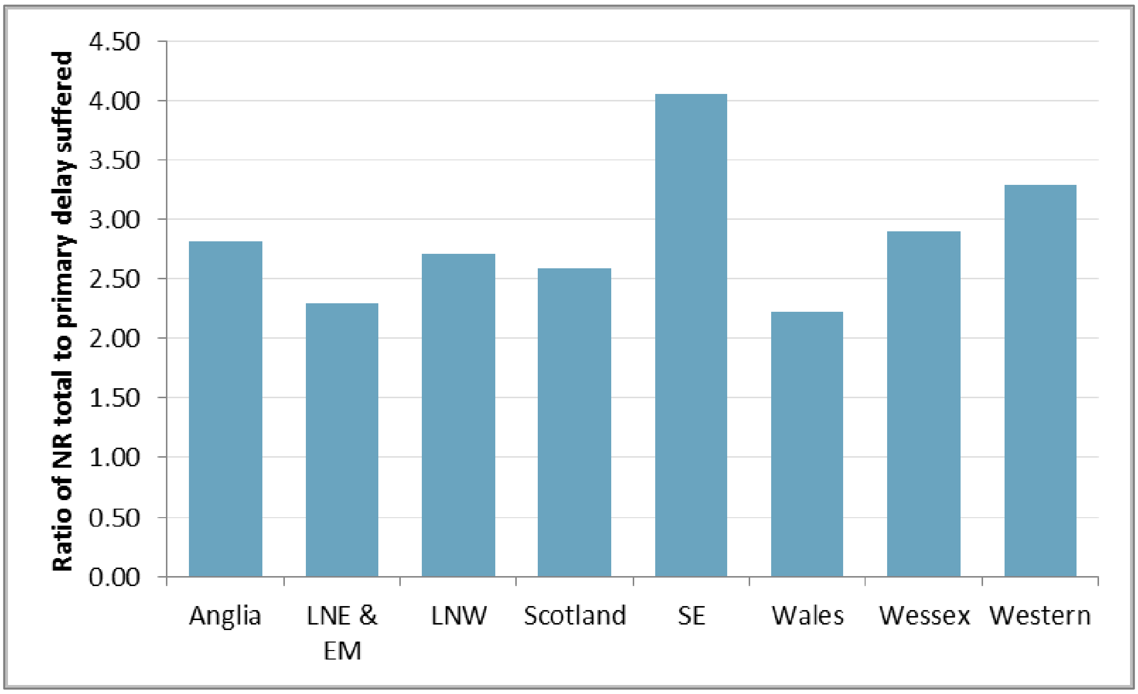
Source: Steer Davies Gleave analysis based on Network Rail data

6.13 The figures suggest very different conclusions about relative levels of performance. In the case of primary MD per train km, the majority of routes have broadly similar levels of performance, with SE exhibiting the worst performance and Scotland performing markedly better than the average. However, if the metric is normalised using service intensity LNE & EM and LNW stand out as the worst performing routes, notwithstanding that the level of train km per track km on these routes is broadly comparable with (if not lower than) that on Anglia, SE and Wessex.

6.14 The use of service intensity as a normalisation factor is arguably less defensible in the case of metrics based on primary delay, since it controls for the tendency for delay to propagate through secondary effects on more intensively used networks. Primary delay may also be affected by service intensity, since the more intensively the infrastructure is used the more likely incidents affecting the train service are to arise. However, in general the relationship will depend on the ratio of total to primary delay across routes. To the extent that this ratio is similar, we would expect the correlation between primary delay and service intensity to be comparable to that illustrated in Figure 4.4 above.

6.15 The figure below shows the ratio of total to primary delay for all routes in 2015/16, and suggests a significant degree of variation. In particular, the ratio for SE is substantially above that exhibited by the other routes, a reflection of the greater tendency for secondary delay to propagate across services operating on SE in view of in view of the greater potential for such services to interact. This suggests that a comparison of routes based on primary delay minutes normalised by service intensity is likely to be misleading, since it will tend to overstate the relative performance of routes with particularly high levels of secondary delay.

Figure 6.8: Ratio of total to primary delay suffered in 2015/16



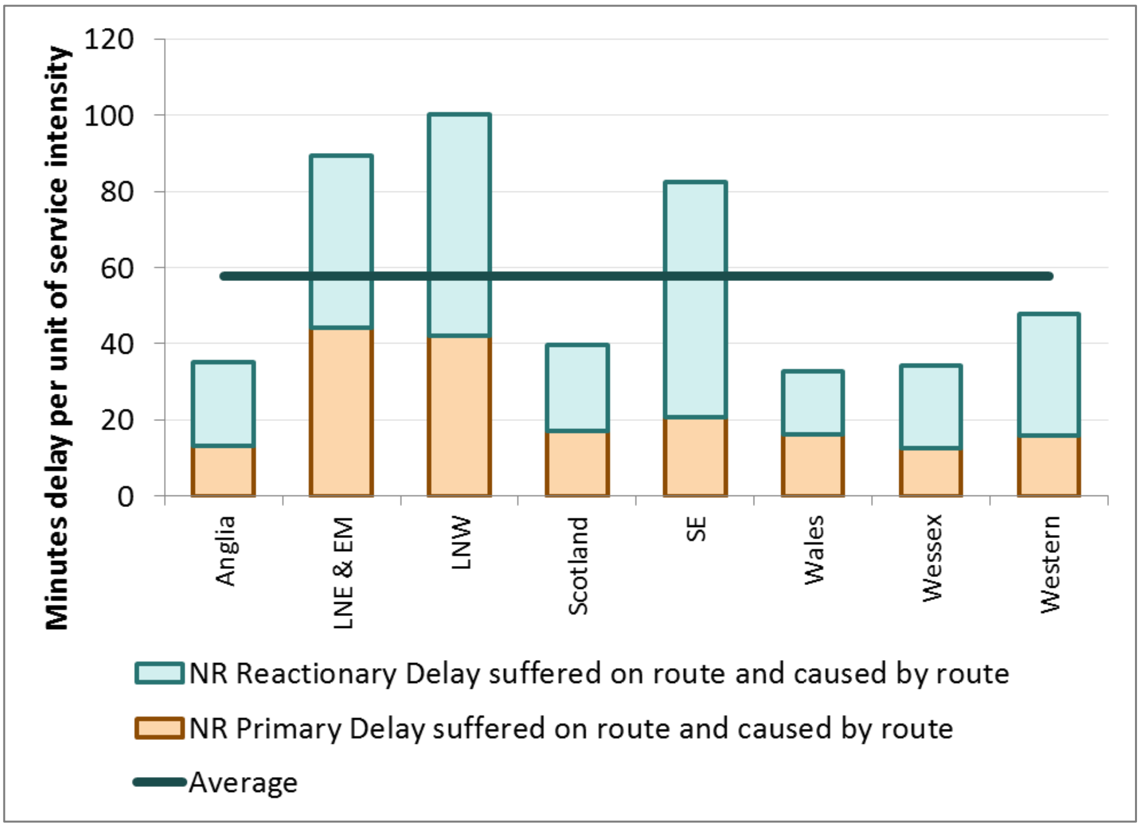
Source: Steer Davies Gleave analysis based on Network Rail data

M4+: Primary and Secondary Minutes Delay caused by and incurred on a given Route

6.16 Perhaps more importantly, Figure 6.8 also demonstrates secondary delay generally accounts for a higher proportion of the total than primary delay, which suggests that a metric based on primary delay alone provides too narrow a view of a route's contribution to performance.

However, as indicated in paragraph 5.13, the metric can be extended to include primary and secondary delay. In the specification provided in the previous chapter, the metric is defined to include all secondary delay that is neither imported nor exported across a route boundary . In effect, this alternative metric captures a much greater proportion of delay attributable to, and managed by, the route than M4, but similarly excludes delay propagated across borders. This means that the application of the normalisation factor is more meaningful, since the train km and track km used to derive the factor are fully aligned with the route geography on which the measured delay actually occurred. The determination of delay included in the metric is illustrated in a worked example provided in Appendix A.

Figure 6.9: Network Rail-caused primary and secondary Minutes Delay per unit of service intensity 2015/16



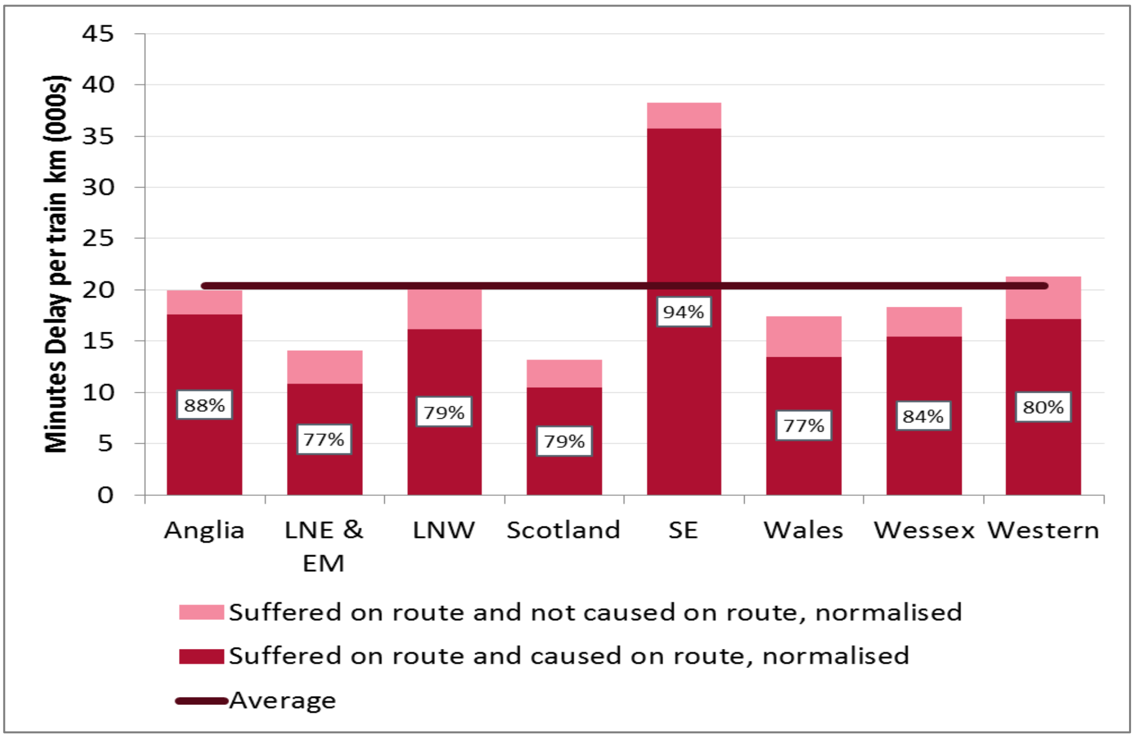
Source: Steer Davies Gleave analysis based on Network Rail data

6.17 Note that we have calculated the extended metric using a normalisation factor based on service intensity, since $M4^{+}$ includes both primary and secondary delay. In practice, these results are similar to those presented in Figure 6.2, and it is therefore open to the objection that it does not add materially to an understanding of relative performance across routes while adding to the complexity of the calculation. However, we suggest that it is preferable to M1 in that it excludes all delay imported across borders, as noted above. At the same time, by including both primary and secondary delay, it is more closely aligned with passenger experience of punctuality.

M5: Secondary Minutes Delay by Route

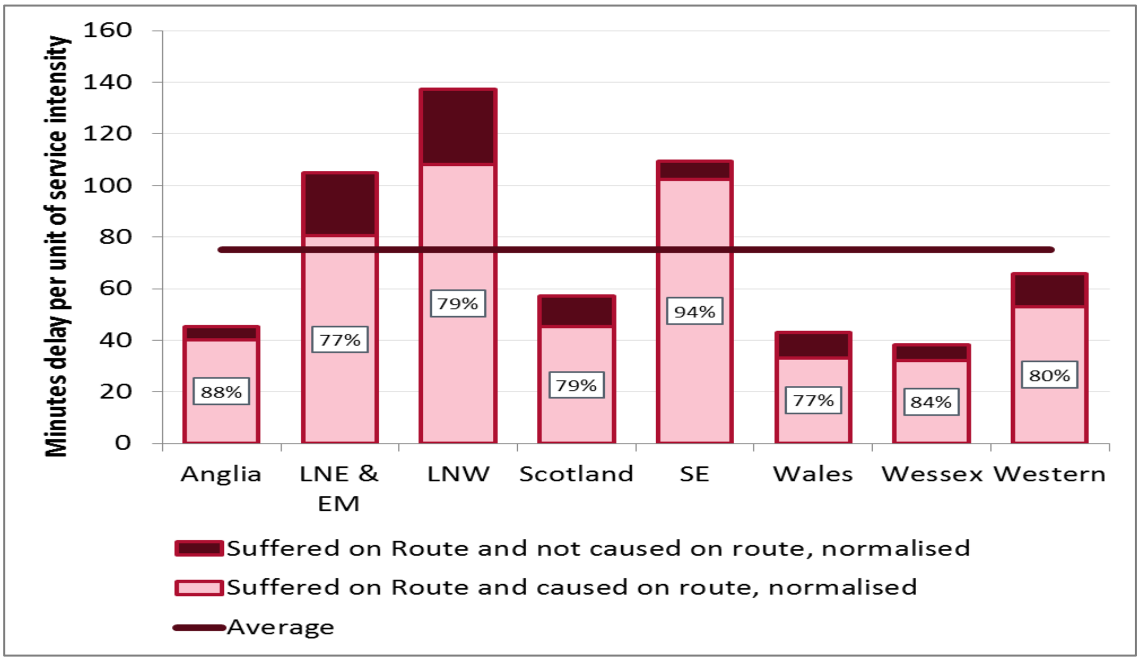
6.18 A comparison of routes in terms of all secondary delay incurred, regardless of cause, is provided in the figures below. These provide an indication of how well different routes manage delay to services whether or not the underlying incident has been caused by them.

Figure 6.10: Secondary Minutes Delay incurred per train km 2015/16



Source: Steer Davies Gleave analysis based on Network Rail data

Figure 6.11: Secondary Minutes Delay incurred per unit of service intensity 2015/16



Source: Steer Davies Gleave analysis based on Network Rail data

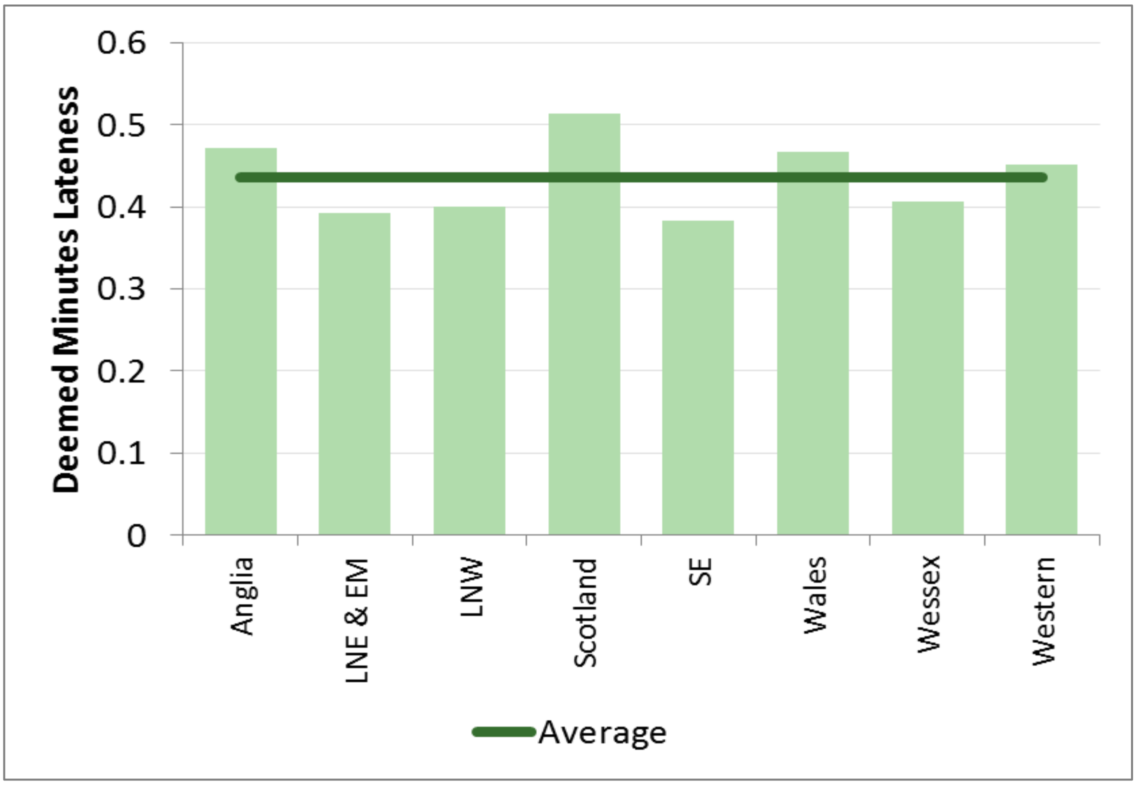
6.19 As in the case of other metrics, the main effect of changing the basis of normalisation concerns the relative performance of SE and the longer distance routes. More specifically, the performance of SE compares more favourably with the other routes when train km per track km is used as the normalisation factor, while that of LNE & EM and LNW compares considerably less favourably. There is also some change to the ranking of other routes, with Wessex emerging as the best performer in Figure 6.11 since intensity of Wessex services is comparable with that on SE.

6.20 In our view, M5 is a useful measure since it captures the outputs from an important element of a route's role, namely management of delay once an incident has occurred. However, as a basis for comparison, it is open to the challenge that it includes delay caused by all parties, at least some which the route may not be able to reduce materially. This includes delay imported from other routes, which, as shown in Figure 6.10, can account for almost a quarter of secondary delay.

M6: Deemed Minutes Lateness by Route

6.21 We have also calculated Deemed Minutes Lateness by route using a similar methodology to that employed in making preliminary calculations of M3 above. As in the case of M3, we emphasise that recording and monitoring systems are not yet able to support an accurate calculation of the metric. Nevertheless, the results shown suggest that it could provide useful supplementary information to the delay-based metrics discussed above, not least because the routes performing relatively well when measured by reference to delay (notably Anglia, Scotland and Wales) appear to be the worst performers when measured in terms of cancellations. We conclude that Network Rail should be encouraged to make the necessary changes to PEARS such that the metric can be introduced as a KPI in the future.

Figure 6.12: Network Rail Deemed Minutes Lateness 2015/16



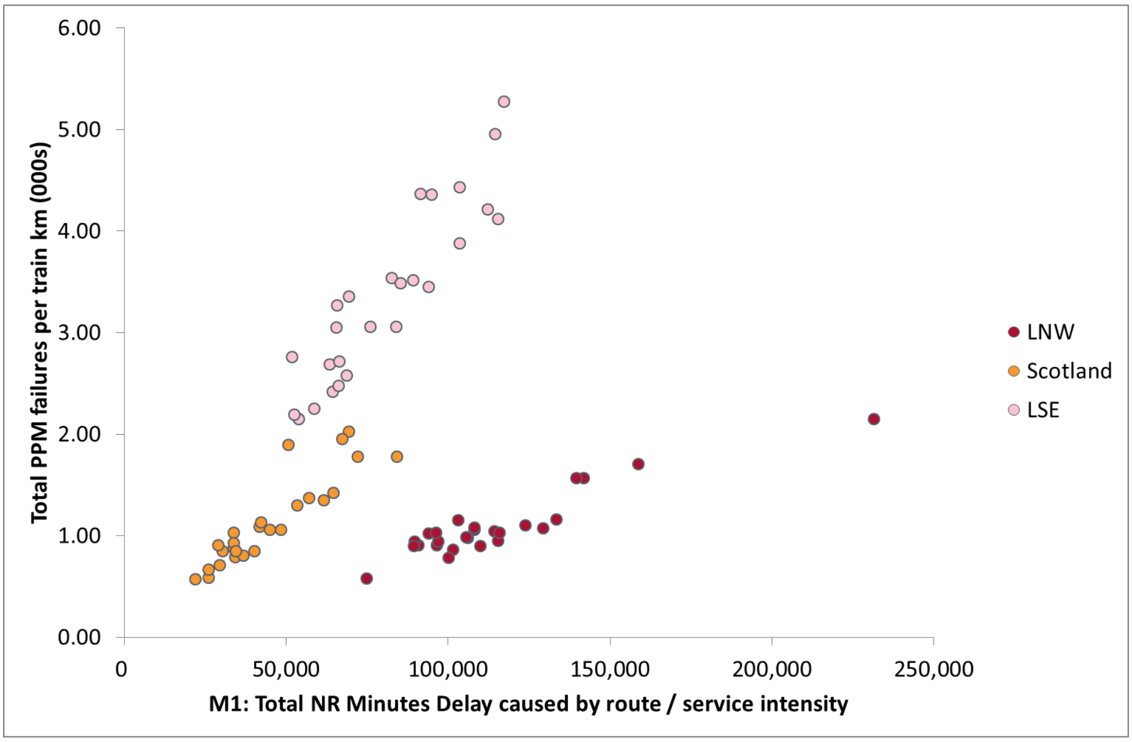
Source: Steer Davies Gleave analysis based on Network Rail data

Relationship with other performance measures

6.22 We have also investigated the relationship between the metrics and key measures included in the route scorecards. As discussed in Chapter 3, it is important that the chosen metric or metrics aligns with the incentives created by the monitoring of other performance measures. Our approach involved investigating the correlation between each metric and two specific measures appearing frequently in the scorecards prepared for CP5, namely PPM and CaSL. In Figure 6.13 we provide an example illustration of the approach for a single metric and scorecard measure combination. For clarity, we have restricted the results to three routes, LNW, SE and Scotland, which collectively cover a wide range of different route characteristics.

6.23 As the figure shows, there is a strong correlation between metric M1 (Network Rail-caused delay normalised using service intensity) and Network Rail-caused PPM failures per train km for all three routes. This can be expressed numerically using a correlation coefficient, which can be calculated for all routes and all combinations of relevant measures. Note that the relationship between any given metric and scorecard measure may vary by route, as the slopes of the different lines implied by the scatter of points in the figure suggests. However, providing the points are tightly grouped around a given line, we can be confident that the metric and measure will move together and that the consistency objective described in Chapter 3 will be met.

Figure 6.13: Relationship between PPM failures and metric M1



Source: Steer Davies Gleave analysis based on Network Rail data

6.24 We have calculated correlation coefficients for 18 combinations of metrics and scorecard measures for each of the eight routes. The results are shown in the tables below. The closer the correlation coefficient is to unity, the stronger the relationship between the metric and measure. We have also colour coded the cells to highlight instances of good and poor correlation (with dark green indicating a coefficient above 0.8 and red indicating one below 0.2).

Table 6.1: Correlation between metrics and PPM failures

Route	M1: Total NR-caused Minutes Delay		M2: Passenger-weighted Delay Minutes	M3: Performance Minutes	M4: Primary Delay		M5: Reactionary Delay		M6: Deemed Minutes Lateness
	Service activity	Service intensity			Service activity	Service intensity	Service activity	Service intensity	
Anglia	0.94	0.94	0.61	0.92	0.86	0.86	0.95	0.95	0.81
LNE&EM	0.96	0.96	0.95	0.90	0.82	0.82	0.96	0.96	0.48
LNW	0.98	0.98	0.87	0.96	0.89	0.89	0.98	0.98	0.87
SE	0.97	0.97	0.15	0.75	0.85	0.85	0.97	0.97	0.60
Scotland	0.96	0.96	0.32	0.86	0.88	0.88	0.98	0.98	0.76
Wales	0.93	0.93	0.49	0.86	0.77	0.77	0.87	0.87	0.64
Wessex	0.96	0.96	0.10	0.90	0.81	0.81	0.91	0.91	0.50
Western	0.94	0.94	0.85	0.81	0.74	0.74	0.95	0.95	0.14

Source: Steer Davies Gleave analysis

Table 6.2: Correlation between metrics and CaSL

Route	M1: Total NR-caused Minutes Delay		M2: Passenger-weighted Delay Minutes	M3: Performance Minutes	M4: Primary Delay		M5: Reactionary Delay		M6: Deemed Minutes Lateness
	Service activity	Service intensity			Service activity	Service intensity	Service activity	Service intensity	
Anglia	0.68	0.68	0.42	0.73	0.73	0.73	0.53	0.53	0.85
LNE&EM	0.61	0.61	0.95	0.80	0.61	0.61	0.60	0.60	0.77
LNW	0.92	0.92	0.86	0.88	0.89	0.89	0.88	0.88	0.92
SE	0.85	0.85	0.71	0.50	0.77	0.77	0.83	0.83	0.56
Scotland	0.73	0.73	0.76	0.95	0.78	0.78	0.72	0.72	0.93
Wales	0.71	0.71	0.84	0.54	0.53	0.53	0.77	0.77	0.60
Wessex	0.79	0.79	0.74	0.44	0.81	0.81	0.71	0.71	0.87
Western	0.71	0.71	0.76	0.49	0.71	0.71	0.65	0.65	0.71

Source: Steer Davies Gleave analysis

6.25 As can be seen from Table 6.2, the majority of the metrics correlate strongly with PPM failures, with M1 and M5 performing particularly well in this respect. We would expect metrics capturing performance of the average train or over the average train km to be closely related to PPM failures per train km, since the chosen normalisation factor is similar. However, the correlation coefficients for metric M2, which introduces passenger weighting into the calculation, and metric M6, which focuses on reliability rather than punctuality, are much lower. The use of M2 in isolation as a basis for comparing routes therefore appears to introduce the risk of creating incentives that run counter to those generated by a key scorecard metric.

6.26 The relatively high level of correlation between a number of metrics and CaSL is perhaps surprising given that, for the most part, they are capturing average punctuality and exclude reliability. However, the fact that M6 is more closely related to CaSL is to be expected since it is a measure of reliability and Deemed Minutes will tend to be correlated with the number of cancellations. These results tend to confirm the value of M6 as a supplementary measure, particularly if metrics such as M3 or M5 were selected as the primary basis for making comparisons of route level performance. More specifically, this would ensure more balanced effort across punctuality and reliability rather than a strong focus on reducing delay as would probably be the result if M4 or M5 were monitored in isolation.

7 Conclusions and recommendations

Appraisal of metrics

Appraisal methodology

7.1 We have undertaken an appraisal of the metrics discussed in the previous chapter against the objectives set out in Chapter 3. This involved considering how far each metric meets the objectives according to the criteria in the table below. While such exercise has inevitably requires judgement, the application of these criteria ensures that the appraisal process is subject to a degree of rigour.

Table 7.1: Appraisal criteria

Objective	Appraisal criteria		
	✓✓✓	✓✓	✓
Comparability	Metric controls for a number of route characteristics and allows meaningful comparison of route contributions to performance	Metric controls to some degree for route characteristics and enables comparison, but a significant risk of challenge if used in isolation	Metric fails to control adequately for route characteristics and any comparison between routes is likely to be challenged if used in isolation
Ease of calculation	Metric can be easily calculated over a number of years using data and systems already available	Metric can be calculated over a number of years using data and systems already available, but limited new processes required	Metric can in principle be calculated over a number of years but would require changes to systems and process that are unlikely to be delivered before CP6
Consistency	Metric clearly correlated with other measures – coefficient of at least 0.6 for the two scorecard measures investigated	Metric does not meet the criteria for '✓✓✓' or '✓'	Metric poorly correlated with other measures – coefficient of 0.2 or below for at least one of the scorecard measures investigated
Customer focus	Calculation includes explicit measure/ parameter reflecting customer experience	Metric clearly correlated with an established measure of customer experience (e.g. lateness)	Metric is poorly correlated with customer experience
Credibility	Metric likely to be understood and supported by a wide range of stakeholders	Metric likely to be understood by key stakeholders and may be accepted following consultation	Use of metric likely to meet with significant resistance/criticism among at least one group of stakeholders

Source: Steer Davies Gleave

Appraisal results

Table 7.2: Route performance metric – appraisal results

Metric	Normalisation	Comparability	Ease of calculation	Consistency	Customer focus	Credibility	Comment
M1	Train km	✓	✓✓✓	✓✓✓	✓✓	✓	Normalisation fails to control adequately for differences between routes. Routes and possibly other stakeholders may well challenge a comparison on the basis of this metric.
M1	Service intensity	✓✓	✓✓✓	✓✓✓	✓✓	✓	Basis of normalisation enables a more meaningful comparison, but still a risk of challenge given the impact of cross-border delay.
M2	Passenger journeys	✓	✓✓	✓	✓✓✓	✓	Weighting by passenger journey makes the comparison less meaningful if the aim is to capture a route's contribution since it cannot influence the distribution of passengers across services. This increases the risk of stakeholder challenge.
M3	Scheduled services	✓	✓	✓✓	✓✓✓	✓✓	Normalisation does not take account of route characteristics and the scope for introducing the metric before CP6 is not clear.
M4	Train km	✓✓	✓✓✓	✓✓	✓	✓✓	A stronger basis for comparison given the focus of route-caused primary delay, but does not capture passenger experience well.
M4	Service intensity	✓	✓✓✓	✓✓	✓	✓	Comparability open to challenge as normalisation factor is inappropriate in the case of primary delay.
M4+	Train km	✓(✓)	✓✓✓	✓✓*	✓✓	✓✓	A compromise between comparability and customer focus – scores reasonably well across all criteria, although may not control adequately for route characteristics.
M4+	Service intensity	✓✓	✓✓✓	✓✓*	✓✓	✓✓	As above, but normalisation improves comparability given inclusion of secondary delay.
M5	Train km	✓	✓✓✓	✓✓	✓	✓	Inclusion of all secondary delay reduces comparability and increases risk of challenge.
M5	Service intensity	✓(✓)	✓✓✓	✓✓	✓(✓)	✓(✓)	As above.
M6	Scheduled services	✓(✓)	✓	✓(✓)	✓✓✓	✓✓	Scores well as a supplementary measure but not as a main metric.

* Not tested explicitly but unlikely to be worse than M4. '✓(✓)' indicates a score between '✓✓' and '✓'.

Scores well against objectives			Scores poorly against objectives

Source: Steer Davies Gleave analysis

7.2 We present the results of the appraisal in the table above. We emphasise that these are based on a limited exercise to apply the criteria described above undertaken within the study team. In our view, they should be subject to further testing and discussion with a range of stakeholders as part of the ORR's ongoing work to define and introduce the new metric.

7.3 Overall M1 and $M4^+$, which are closely related, are the best performing metrics, particularly if normalised using a measure of service intensity. Both can be calculated relatively easily and move in a way that is consistent with the scorecard measures investigated. M4 is more comparable, as it excludes secondary delay imported across borders and would therefore be subject to less risk of challenge. The other metrics are subject to a number of limitations and in our view could not be introduced as single regulatory metrics. M3 and M6 are also ruled out as the ability to introduce them before CP6 is highly uncertain.

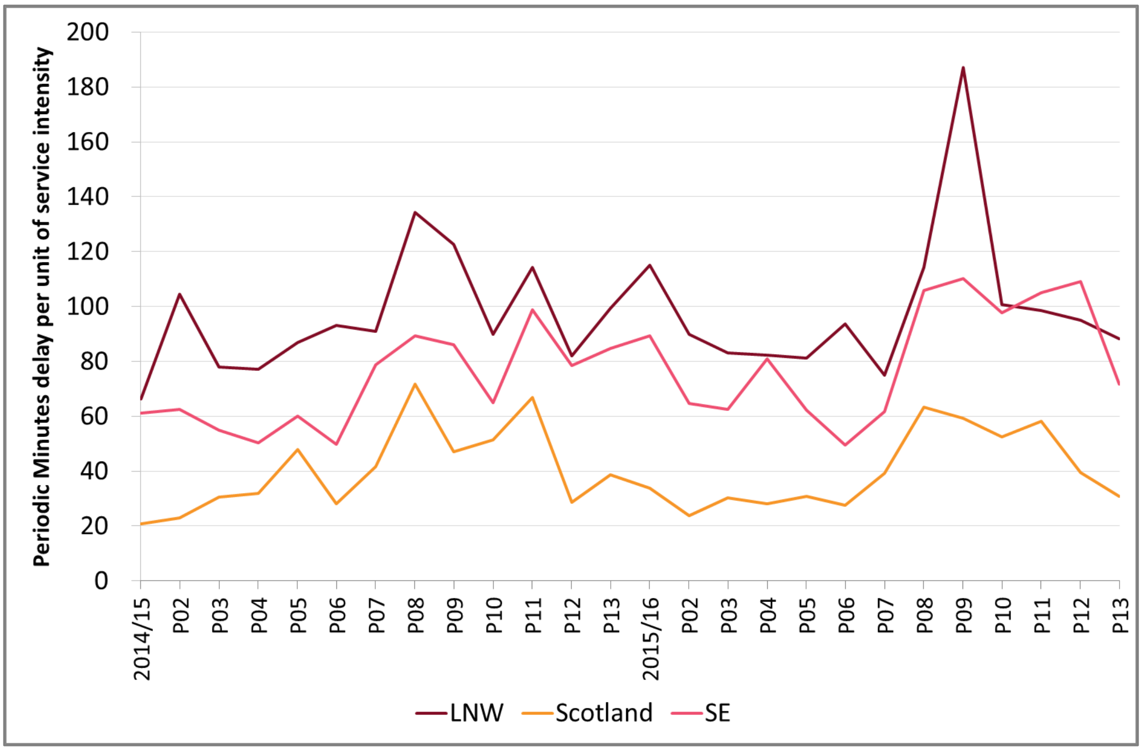
Recommendations

7.4 In the light of these results, we recommend that the primary metric used for comparing route level performance should be defined as:

$M4^+$- The sum of Network Rail-caused primary and secondary delay caused by, and suffered on, a given route, normalised using train km per track km.

7.5 The trend in this metric during 2014/15 and 2015/16 for three routes is shown in Figure 7.1. This indicates broadly stable performance, although normalised delay appears to have been increasing on all three. As already noted, we understand that sufficient data is available to enable the investigation of the metric since 2009, and this would allow the calculation of a moving annual average in order to assess trends in performance more easily.

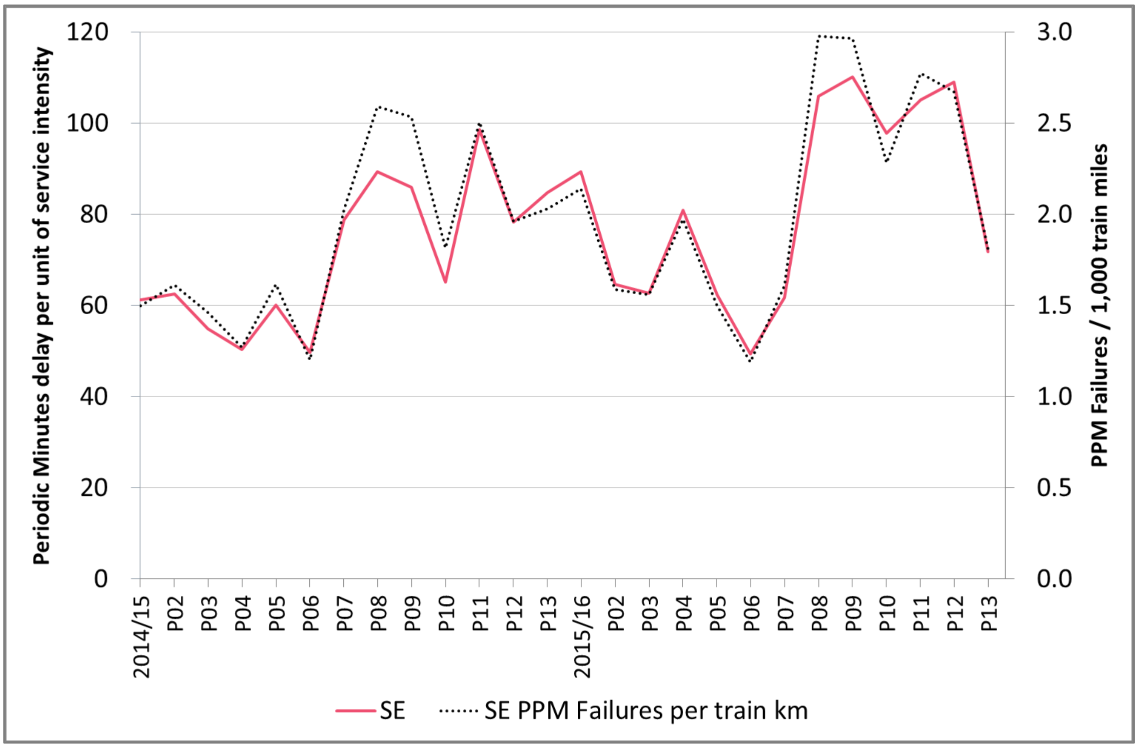
Figure 7.1: $M4^+$ by period in 2014/15



Source: Steer Davies Gleave analysis based on Network Rail data

7.6 We have also investigated the relationship between the metric and PPM, a performance measure included in a number of the initial route score cards that we have reviewed. Figure 7.2 demonstrates a high degree of correlation for the SE route, providing assurance that M4⁺ will support and reinforce other route performance incentives.

Figure 7.2: M4⁺ per period and PPM failures



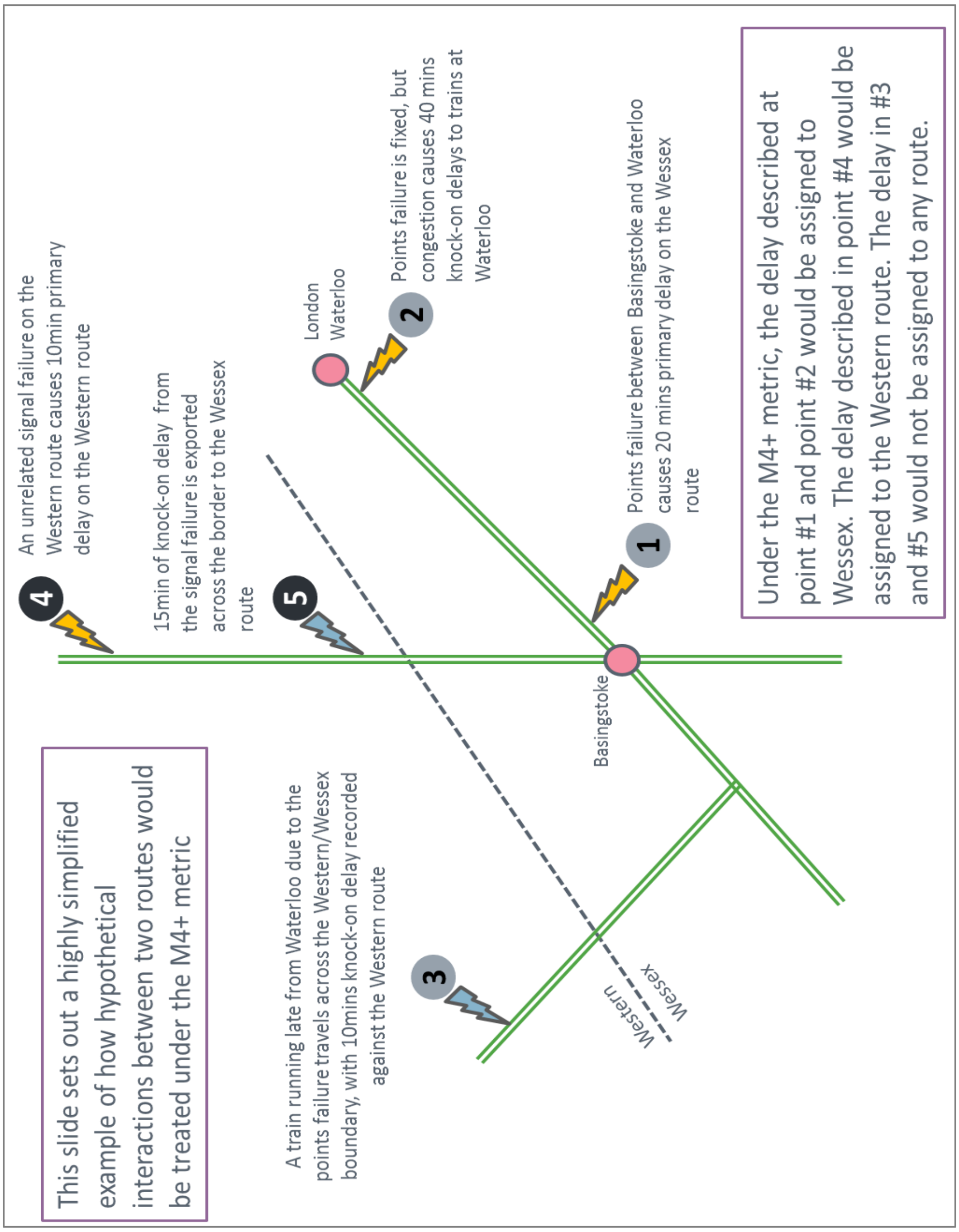
Source: Steer Davies Gleave analysis based on Network Rail data

7.7 In our view, M4⁺ could be used as a formal regulatory metric, with routes reporting against targets set at a periodic review. However, we consider that it should be supplemented by other metrics, monitored on a less formal basis but nevertheless used to inform ongoing review and challenge of route level performance as part of a broader regulatory engagement process. Our proposed supplementary metrics are as follows:

- *M4 - The sum of Network Rail-caused primary delay caused by, and suffered on, a given route, normalised using train km;*
- *M5 – The sum of all secondary delay suffered on a route, regardless of cause, normalised using train km per track km; and*
- *(At a future date following the necessary changes to existing systems) M6 – Average deemed minutes attributable to a route, derived using parameters and metrics already used in the calculation of Schedule 8 payment.*

7.8 Together, these metrics provide a basis for monitoring different aspects of a route's contribution to performance and capture passengers' experience of both punctuality and reliability. M4⁺, M4 and M5 could all be introduced before CP6 and readily analysed to identify historical trends. We suggest that Network Rail is encouraged to undertake the necessary modifications to PEARS in order to support the calculation of M6 as soon as possible.

Appendix A: Application of Metric M4+



CONTROL INFORMATION

Prepared by	Prepared for
Steer Davies Gleave 28-32 Upper Ground London SE1 9PD +44 20 7910 5000 www.steerdaviesgleave.com	Office of Rail and Road One Kemble Street London WC2B 4AN

SDG project/proposal number	Client contract/project number
23095701	ORR CT 16-81

Author/originator	Reviewer/approver
Steer Davies Gleave	David Ford

Other contributors	Distribution
Simon Ellis, David Ford, Jonathan Platten	*Client:* ORR *SDG*: Project file

Version control/issue number	Date
V0.9 final	31 March 2017

DNV UKAS QUALITY MANAGEMENT 013