Joel Strange
Kings Place
90 York Way
London
N1 9AG
T 020 3356 9319
Joel.strange@networkrail.co.uk

23 August 2013

Dear colleague

Schedule 8 Network Rail benchmarks in Control Period 5

ORR is currently undertaking its periodic review (PR13), which will set the outputs and funding for Network Rail for Control Period 5 (CP5). As part of this process, the industry is reviewing the parameters of the Schedule 8 regime, which provides compensation for unplanned disruption.

Central to the successful functioning of the Schedule 8 regime are the 'benchmarks' or 'performance points', which provide the 'targets' in the Schedule 8 regime. Network Rail's benchmarks are calibrated so that the Schedule 8 regime is financially neutral if it meets its regulatory performance targets, defined in terms of the Passenger Performance Measure (PPM) and Cancellations and Significant Lateness (CaSL). The intention of Schedule 8 – and the level of the benchmarks in particular – is that if Network Rail outperforms its regulatory targets, it receives a net payment from TOCs in light of the increased farebox revenues they enjoy resulting from this 'above target' performance. Conversely, if Network Rail underperforms its targets, Schedule 8 is designed so that Network Rail makes payments to operators to reflect passenger revenues being lower than they otherwise would be.

I would like to take the opportunity to thank you for your contributions to date in developing benchmarks for Schedule 8 for CP5. In particular, TOCs and Network Rail routes have made a substantial contribution towards developing a benchmarking methodology that is a significant improvement on the approach pursued previously.

Purpose of this letter

The purpose of this letter is to consult with operators and other stakeholders on the assumptions and detailed methodology employed by Network Rail to arrive at Schedule 8 benchmarks for CP5.

The letter also sets out a set of Schedule 8 benchmarks. These benchmarks are consistent with ORR's draft determination on CP5 performance[1]. We would emphasise, however, that we **do not believe that the PPM performance trajectory set out in the draft determination is the most realistic assessment of likely CP5 performance, particularly in the first years of the control period. We expect the final set of Schedule 8 benchmarks to be updated once the CP5 regulatory trajectory is finalised as part of ORR's final determination.**

This consultation relates only to Network Rail Schedule 8 benchmarks in the passenger regime – a separate process is being undertaken to set benchmarks for the freight and charter regimes for CP5. The deadline for responses to this consultation is **20 September 2013. Ultimately, any decision on Schedule 8 benchmarks for CP5 will be an issue for ORR.**

Previous consultation on methodology and Schedule 8 principles benchmarks

In April 2013, we wrote to the industry to gather views on a methodology and set of principles for setting Schedule 8 benchmarks for CP5[2]. We received a number of responses to our consultation, and these have heavily informed our work. A summary of the responses and how these have informed our thinking is contained in our conclusions document on Schedule 8 benchmarks[3]. Important features of our approach that have been influenced by stakeholders' comments include using more granular regression analysis, and estimating separate relationships for the punctuality and cancellations elements of performance.

The principles that we proposed for setting Schedule 8 benchmarks gained widespread support. We proposed these principles to ORR, and it has recently decided the set of principles on which Schedule 8 benchmarks should be set for CP5. ORR's principles are very similar to Network Rail's proposals, albeit with some minor amendments. ORR's principles are as follows:

i. For each year of CP5, Schedule 8 Network Rail benchmarks should be consistent with achieving the annual performance targets specified in our draft determination, such as PPM and CaSL. If output targets are changed for the final determination, the benchmarks will need recalculating on time for the beginning of CP5;
ii. Subject to i., CP5 Schedule 8 Network Rail benchmarks should reflect industry's view on expected CP5 performance by TOC, and therefore be consistent with whole CP5 performance (e.g. PPM and CaSL) trajectories at train operator level, which should be developed by Network Rail working with TOCs;

[1] In light of ongoing JPIP and other discussions with a number of TOCs, particularly long distance ones, we have assumed that the PPM MAA performance of all TOCs will be at least at the 88 percent level by the end of CP5.
[2] http://www.networkrail.co.uk/using-our-network/on-train-metering/schedule-8-benchmarks.pdf
[3] http://www.networkrail.co.uk/Schedule8benchmarksConclusionsLetter.pdf/

iii. Schedule 8 Network Rail benchmarks should be set on the basis of the most recent data available at the time of calculation and relationships between Schedule 8 average minutes lateness (AML) and the performance targets specified in our draft determination;
iv. A rebenchmarking exercise should take place if we think there are material changes to timetables, for example as a result of refranchising. These new benchmarks will be active from the date of the material change to the timetable or the proposal for a change in the benchmark, whichever occurs later; and
v. If 'change control' is used in CP5 to adjust regulatory outputs, appropriate adjustments should also be applied to Schedule 8 Network Rail benchmarks. The new benchmarks will be active from the date following the adjustment to the regulatory outputs.

Schedule 8 workstreams

The Schedule 8 benchmarking exercise involves three principal workstreams:
i. Establishing 'baseline' performance;
ii. Establishing TOC-level PPM/CaSL performance trajectories; and
iii. Converting TOC-level PPM/CaSL trajectories to Schedule 8 benchmarks.

Each of these workstreams is discussed, below, and we ask a number of specific consultation questions.

Establishing 'baseline' performance

The first workstream involves establishing the historic average performance, as measured by 'performance minutes' (the measure of performance used in Schedule 8). This is done for each Service Group using a two-year 'calibration period', typically taken to be the financial years 2010-11 and 2011-12 (although it has been appropriate to deviate from this time period in some circumstances, for example if there was a fundamental change of services). This process was led by consultancy Halcrow, which has had a comprehensive engagement programme with TOCs. This workstream came to an end recently. The baseline figures used in this analysis were provided by Halcrow in mid-July 2013. These baselines are repeated in Appendix 1 (Appendix 1 is specific to each TOC). **We understand from Halcrow that further changes may take place for a small number of TOCs, and further changes made by Halcrow will need to be reflected in final benchmarks.**

Question 1
Do you have any comments on the baseline performance figures provided by Halcrow, and reproduced in Appendix 1?

Establishing TOC-level PPM/CaSL performance trajectories

Network Rail route performance teams have worked with TOCs to produce PPM and CaSL trajectories for each year of CP5, for each TOC. The performance trajectories used for benchmarking purposes are set out in Appendix 1.

These come slightly short of 'adding up' to ORR's draft determination performance trajectory. And, as noted above, our current view is that the likelihood of ORR's draft determination trajectory being achieved in the first years of CP5 is low. As described in detail below, we have made adjustments to the benchmarks shared as part of this consultation to ensure that they are draft determination compliant in aggregate.

Question 2
Do you have any comments on the TOC-specific PPM and CaSL trajectories set out in Appendix 1?

Converting TOC-level PPM/CaSL trajectories to Schedule 8 benchmarks

A core principle is that Schedule 8 benchmarks should be calibrated so as to be consistent with regulatory targets (i.e. Schedule 8 should be financially neutral to Network Rail if it achieves these targets). The regime is designed so that train operators make payments to Network Rail for outperformance of these targets, and that Network Rail compensates train operators for underperformance of targets.

Network Rail's regulatory performance targets are expressed in terms of PPM and CaSL. On the other hand, Schedule 8 uses a different measure of performance – Average Minutes Lateness (AML) and Deemed Minutes Lateness (DML). It is therefore necessary to 'convert' the PPM/CaSL trajectories into the 'language' of Schedule 8. This workstream has been led by Network Rail's central Regulatory Economics and Performance teams, and has involved consultation with both TOCs and Network Rail routes.

A detailed methodology is set out in Appendix 2 to this letter. At a high level, the methodology consists of the following steps:

- Regression analysis of historic regulatory performance measures (PPM, CaSL and, by implication, punctuality) and Schedule 8 performance measures (AML and DML) to estimate the relevant relationships; and
- Using these relationships to 'translate' changes in regulatory performance measures to the Halcrow baselines, in order to arrive at a Schedule 8 benchmark by Service Group.
- Applying 'shares' to recognise which parties (Network Rail, the TOC or both (joint schemes) are assumed to 'deliver' the changes in PPM/CaSL.

We have prepared a model for each TOC, and these models have been shared with operators as part of this consultation. The models have been reviewed by

consultants Steer Davies Gleave, in order to ensure that they are fit for purpose and computationally correct. Steer Davies Gleave's report is contained in Appendix 3. Checks are continuing, and we will be in touch with parties if any anomalies are identified. We would invite operators to let us know at the earliest possible stage, should they identify any areas of concern when using the models.

Question 3
Do you agree with the methodology set out in Appendix 2? If you do not agree, please explain why not, and provide **specific suggestions** around ways in which the methodology can be improved.

As part of our informal consultation process in advance of publishing this letter, we spoke to all TOCs in order to discuss the relationships between performance minutes and regulatory performance measures, and gather information to strengthen these relationships. We found these meetings to be extremely informative and we are grateful to operators for spending the time to meet with us – we believe that the robustness of the work has been strengthened significantly by operators' contribution.

Three particular issues were discussed with TOCs:

- Whether the assumption that the relationship between performance minutes and regulatory performance measures is linear is appropriate;
- Whether there have been any important 'structural breaks' that may have impacted the relationships (e.g. major timetable recasts or the introduction of new rolling stock); and
- Whether there have been certain genuinely 'unusual' periods which may skew the relationships which should be omitted from the analysis (e.g. the Olympics period).

Any specific changes made in relation to these areas are set out in Appendix 1. In all cases, we have only made changes where there are both operational and statistical grounds to do so.

Question 4
Do you have any comments on the TOC-specific assumptions made in relation to the regression analysis set out in Appendix 1?

Other TOC-specific issues

There are a number of TOC-specific issues and assumptions that have come out of this process. For example, for some TOCs, a different 'baseline' period was used to reflect the introduction of new services or deep-rooted changes to timetables. Any TOC-specific issues are set out in Appendix 1.

Question 5
Do you have any comments on the other TOC-specific issues set out in Appendix 1? Do you agree with the assumptions about the 'shares' assumed to deliver performance improvements?

We are aware of a small number of circumstances where material changes to timetables will take place following the benchmark period (since the beginning of 2012-13) and before the end of CP4 i.e. in the final two years of CP4. It will be necessary to make adjustments to benchmarks as a result of these changes prior to the beginning of CP5, or during CP5 itself. Work around this will need to be taken forward locally, by means of discussions between TOCs and Network Rail routes.

Question 6
Are you aware of any changes to timetables or other changes that will take place in the final two years of CP4 that will materially affect the level of performance minutes or the relationship between performance minutes and regulatory measures?

Network Rail benchmarks for CP5

The primary output of this exercise is a set of Schedule 8 benchmarks for each Service Group for every year of CP5. Benchmarks are set out in Appendix 4 to this consultation. Two sets of benchmarks are shown for each Service Group in Appendix 4. The first set is consistent with the PPM/CaSL trajectories shared between TOCs and Network Rail routes during the summer. However, when aggregated, these come slightly short of fulfilling the draft determination trajectory. We have, therefore, made further adjustments to the Network Rail benchmarks in order to ensure that the benchmarks are consistent with the draft determination in aggregate[4]. As such, the second set of benchmarks is consistent with the draft determination trajectory. It should be emphasised that these adjustments are **only** made to the Network Rail Schedule 8 benchmarks and are not a commitment for further performance improvement or an alteration to any agreed targets.

The benchmarks contained in Appendix 4 are consistent with ORR's draft determination on CP5 performance. We would emphasise, however, that we **do not believe that the PPM performance trajectory set out in the draft determination is the most realistic assessment of likely CP5 performance, particularly in the first years of the control period. We expect the final set of Schedule 8 benchmarks to be updated once the CP5 regulatory trajectory is finalised as part of ORR's final determination.**

[4] At present we are unsure around the specific schemes and activities that will 'fill the gap', and therefore which operators will benefit from better performance. The planning process continues. In order to appropriately reflect this uncertainty, a national overlay has been added to the Network Rail benchmarks. We have assumed a consistent overlay on all operators equivalent to less than 0.02 percent PPM, plus an additional overlay based loosely on current performance and the PPM required by 2016/17 (by which time, the trajectories shared between TOCs and routes over the summer 'add up' to the draft determination trajectory). It should be emphasised that these adjustments are only made to the Network Rail Schedule 8 benchmarks and are not a commitment for further performance improvement or an alteration to any agreed targets.

In addition, we note that the statistical distribution of performance and Schedule 8 is 'asymmetric', in the sense that there is greater risk on the 'downside' than 'upside'. A particular source of asymmetry is that payment rates and/or benchmarks could be reopened – for purposes of developing Schedule 8 benchmarks and performance plans, we have assumed that there will be no reopeners in CP5 (except for 'usual' reasons such as re-writing of timetables). This means that Schedule 8 benchmarks should be set on the basis of a 'P-mean' rather than a P-50 in order to maintain financial neutrality of the Schedule 8 regime. Since the regulatory performance trajectory published in the draft determination is set on a P-50 basis, the final set of benchmarks will need to be adjusted to a P-mean level to ensure expected financial neutrality of the regime.

If ORR decides that benchmarks are to be set on a P-50 rather than 'expected' basis, it will be important that the additional costs are reflected in Network Rail's efficient expenditure projections for CP5.

Why are benchmarks changing?

For a number of Service Groups and TOCs, the proposed Network Rail benchmarks are noticeably different to those currently in place. This is a common occurrence at the beginning of a new control period as the performance regime is reset. There are a number of reasons for these changes, which we describe, below.

Updated regulatory trajectory

The principal reason for the changes is that the regulatory performance trajectory contained in ORR's draft determination assumes PPM of 92.2 percent in the first year of CP5, compared to 92.6 percent in the final year of CP4 for England & Wales. In order to maintain Schedule 8 at a level which is financially neutral if regulatory trajectories are met, Schedule 8 benchmarks are required to change.

Resetting Service Group level trajectories

Performance at individual Service Group level often changes significantly over the course of the control period, regularly becoming significantly out of kilter with the trajectory assumed as part of setting Schedule 8 benchmarks. In CP4 for example, this issue has arisen because a generic trajectory was applied to each TOC, whilst individual Service Group changes have often been very different to the 'TOC average'. These changes will be reflected through movements in the benchmarks as a result of the recalibration.

Disagreement between 'theory' and 'practice' in CP4

When the operation of Schedule 8 is examined 'in the round' rather than on the basis of individual TOCs, the empirical evidence suggests that there may be some difference between the 'theory' and 'practice' of Schedule 8 for CP4. In particular, whilst Schedule 8 should be financially neutral to Network Rail when it reaches its

regulatory targets (and it should involve a net payment to Network Rail when it exceeds those targets), this does not appear to have been borne out in practice during CP4.

For example, in the first year of CP4, 91.5 percent PPM was delivered nationally, compared to a target of 91 percent. In contrast, Network Rail's Schedule 8 income was negligible (£2m in 2009/2010 prices). Similarly, whilst it is inevitable 'mismatches' will arise in Schedule 8 benchmarks in some cases – in the sense that outperformance is accompanied by a financial loss or underperformances is accompanied by a financial gain – there is ex-post evidence that there is a systematic tendency for such mismatches to have involved a net loss to Network Rail in CP4. In CP4 to date, at the sector level (at which regulatory trajectories are set in CP4), there have been ten cases whereby Network Rail has 'lost' financially in periods of outperformance for every one instance in which it has 'gained' financially when targets have been missed.

We believe that the improved methodology deployed for CP5, which has been developed with the industry working closely in collaboration, has helped remove this systematic 'mismatch' for CP5 benchmarks. Overall, whilst there is no contractual mechanism available to seek a refund of this money from payments made in CP4, it is important to correct any systematic payments moving forward so as to deliver value for money for the funders of the railway.

Workshop and next steps

We will hold a workshop at 15:00 on 6 September 2013, which will give TOCs and other stakeholders the opportunity to ask questions and provide feedback to Network Rail on any of the issues raised in this letter, or on Schedule 8 benchmarks in CP5 more generally. If you would like to attend this workshop, please contact Elyse Stoten (Elyse.Stoten@networkrail.co.uk).

The deadline for responses to this consultation is **20 September 2013.**

Following receipt of consultation responses, we will evaluate responses made and make changes where appropriate. We will then make a – potentially revised – proposal to ORR. We expect these submissions to be on a 'rolling basis', beginning in early October. We will, of course, share our communications with ORR in relation to TOC benchmarks with the relevant TOC on a case-by-case basis. **Ultimately, any decision on Schedule 8 benchmarks for CP5 will be an issue for ORR.** Working with ORR, we are planning to finalise all benchmarks in time for ORR's final determination on 31 October. However, given that timescales have been pushed back to accommodate further discussions between TOCs and Network Rail routes around appropriate performance trajectories, it is possible that some benchmarks will be finalised shortly after publication of the final determination. As noted above, the final set of benchmarks will need to be consistent with the performance trajectory contained in the final determinations. We will, of course, provide updates as appropriate.

Given the TOC-specific nature of this consultation, we do not plan on publishing responses on our website. However, we do intend sharing responses with ORR. Please state in your response if you **do not** want it to be shared with ORR.

If you have any questions in the meantime, please contact me using the details above.

Yours faithfully

Joel Strange
Senior Regulatory Economist

Appendix 1

TOC-specific assumptions

TOC-specific

Appendix 2

Detailed methodology

Introduction

As described in our consultation letter of May 2013, the exercise to establish Schedule 8 benchmarks for CP5 consists of three principal workstreams:

i) Establish 'baseline' performance – This workstream involves establishing historic average performance, as measured by 'average minutes lateness' (the measure of performance used in Schedule 8). This is done for each Service Group using a two-year 'calibration period', typically taken to be the financial years 2010-11 and 2011-12, as determined by ORR (although it has been appropriate to deviate from this time period in some circumstances). This process has been led by consultancy Halcrow, which has had a full engagement programme with TOCs.
ii) Establish TOC-level PPM/CaSL performance trajectories – Network Rail route performance teams have worked with TOCs to produce PPM/CaSL trajectories for each year of CP5, for each TOC. The versions used for the benchmark-setting exercise are, in effect, consistent with the draft determination. However, as set out in the body of the letter, the ORR trajectory is unlikely to be realistic.
iii) Convert TOC-level PPM/CaSL trajectories to Schedule 8 benchmarks – This workstream has been led by Network Rail's central Regulatory Economics and Performance teams, and has involved consultation with both TOCs and Network Rail routes.

The purpose of this appendix is to set out the methodology used for (iii), above i.e. to translate regulatory performance trajectories – and in particular the PPM and CaSL trajectories at TOC level – into Schedule 8 benchmarks. This methodology is then applied in a set of spreadsheet models which have been developed by Network Rail. These spreadsheet models have been reviewed by consultants Steer Davies Gleave (SDG). The SDG report is in Appendix 3. The output of these models is a draft determination consistent set of Schedule 8 benchmarks for CP5. These possible benchmarks are set out in Appendix 4 to this letter. One model has been prepared in relation to each TOC, and we have shared the models with TOCs as part of this consultation exercise.

The next section describes the purposes and structure of the models. We then describe in detail 'how' the models serve their purposes of estimating the relevant relationships and then applying these to derive Schedule 8 benchmarks for CP5.

Model purpose and structure

The purposes of the models are to:

i) Estimate the relationships between the Schedule 8 measure of performance ('average' and 'deemed' minutes lateness, or AML and DML) and measures of performance used for industry planning and regulatory targets (PPM and CaSL); and
ii) Apply the relationships in order to 'translate' PPM and CaSL performance trajectories into Schedule 8 benchmarks defined in terms of AML and DML.

The models have been developed in Microsoft Excel. Whilst Excel can be cumbersome and imposes certain constraints on the type of statistical methods that can be used, we have used the software so that results can be readily shared and replicated by ORR, TOCs and other stakeholders.

The models contain a set of inputs, including historic data on various relevant measures of performance, as well as projected future performance trajectories (in terms of PPM and CaSL, as in workstream (ii), above). The models contain a number of calculation sheets. These sheets perform various transformations of the data and undertake the regression analysis (as described below). The principal model outputs are the results of the regression analyses and the calculated Schedule 8 benchmarks.

The relationship between Schedule 8 performance minutes and regulatory performance measures

In order to establish the appropriate relationships between Schedule 8 performance minutes and regulatory output measures we use regression analysis applied to historic data.

Data

As default, we have used all data in the first four years of CP4 to date to undertake this analysis. The data is periodical. As explained below, however, the modelling has been designed so that it can accommodate 'structural' changes in the relationships, for example if a major timetable recast took place within CP4, and can be set to 'omit' certain periods of data, if appropriate.

There are two principal inputs used in the models.

1. Performance minutes - AML, DML and total performance minutes (AML + DML)

   DML (deemed minutes lateness) and total performance minutes have been extracted from PEARS (Paladin Data Extract and Reporting System), which produces analysis reports focusing on train performance and delay attribution. This contains a performance minutes measure for each period and service group from 2001/02 period 1 to 2013/14 period 1. AML (average minutes lateness) is then calculated from these two data inputs.

2. PPM, Cancellations and punctuality

   PPM data has been taken directly from TRUST (Train Running System, TOPS – Total Operators Processing System), which records details of train operational data as compared with schedule. This data source contains an entry for each period and service group from 2004/05 period 1 to 2013/14 period 2. This also contains metrics to allow for the calculation of punctuality, Cancellations, and average PPM measures of a specified period of time (Train Count, PPM Passes, CaSL Failures and All Cancellations). In the model, we have calculated average PPM, Cancellations and punctuality for the recalibration period.

These two data sources are widely-used through the industry, and are generally accepted as the most accurate sources of performance-related data.

The data required has been extracted straight from the data sources explained above, and placed into the model for each TOC. This data has then been manipulated within the model in order to perform the regression analysis. No changes or transformations have been made to the raw data, unless stated otherwise.

Methodology

Before undertaking the analysis, many discussions were had as to how to relate regulatory targets (PPM and CaSL) to AML / DML. Two options were considered:

i) Estimate the relationship between the regulatory targets and delay minutes, and also estimate the relationship between delay minutes and AML / DML. These two relationships would then be used to relate PPM and CaSL trajectories to AML / DML benchmarks.
ii) Estimate the relationship between the regulatory targets and AML / DML directly.

This was discussed in detail at the Schedules 4 & 8 industry groups, and following careful consideration of the issues, the industry decided that it would be more appropriate to pursue option two, whereby the relationship is estimated directly. ORR has confirmed that this option is preferable in its set of principles for calculating the Schedule 8 Network Rail benchmarks for CP5. Specifically, it states that:

"Schedule 8 Network Rail benchmarks should be set on the basis of the most recent data and relationships, available at the time of calculation, between Schedule 8 average minutes lateness (AML) and the performance targets specified in our draft determination"

Crucially, and in contrast to CP4, there are no formal delay minutes targets to 'drive' the relationship for CP5. As such, the second option is likely to minimise errors, as it will involve only one set of estimations. Determining the relationship between PPM / CaSL and AML / DML via delay minutes will incorporate additional risk, as two sets of errors will be captured. Moreover, the industry's emphasis is increasingly moving away from targeting delay minutes towards measures of performance that passengers really care about, such as PPM and lateness.

Another important issue is the determination of the appropriate variables to use in the regression analysis.

During our May 2013 consultation, we received feedback from a number of stakeholders suggesting that we should seek to model average and deemed minutes lateness separately, as opposed to deriving an 'aggregate' relationship between total performance minutes (i.e. AML plus DML) and PPM. As such, we have sought to model the following relationships at Service Group level by means of regression analysis (Ordinary Least Squares (OLS)):

$$AML_{it} = \beta_i^0 + \beta_i^1 Punc_{it} + \beta_i^2 {Punc_{it}}^2 + \beta_i^3 Punc_{it} * D_{it} + \beta_i^4 D_{it} + e_{it} \quad (1)$$

$$DML_{it} = \gamma_i^0 + \gamma_i^1 Canc_{it} + \gamma_i^2 Canc_{it}^2 + \gamma_i^3 Canc_{it} * D_{it} + \gamma_i^4 D_{it} + v_{it} \quad (2)$$

$$TML_{it} = \phi_i^0 + \phi_i^1 PPM_{it} + \phi_i^2 PPM_{it}{}^2 + \phi_i^3 PPM_{it} * D_{it} + \phi_i^4 D_{it} + u_{it} \quad (3)$$

where for Service Group i at time period t:

- AML_{it} is total average minutes lateness (i.e. for both Network Rail and the TOC);
- $Punc_{it}$ is 'punctuality', defined as PPM passes plus cancellations, divided by train count;
- DML_{it} is total deemed minutes lateness (again, the sum of both Network Rail and TOC minutes);
- $Canc_{it}$ is cancellations divided by train count;
- TML_{it} is 'total' delay minutes for Network Rail and the TOC, i.e. $TML_{it} = AML_{it} + DML_{it}$;
- PPM_{it} is PPM, defined as the number of PPM passes over train count;
- D_{it} is a 'dummy' or 'binary' variable, taking the value 0 before a 'structural break' (see below), and 1 afterwards; and
- e_{it}, v_{it} and u_{it} are error terms.

The βs, γs and ϕs are parameters to be estimated. Note from the definition above that the punctuality measure can be calculated as:

$$Punc_{it} = PPM_{it} + Canc_{it} \quad (4)$$

As PPM is defined as $\frac{PPM\ passes}{Train\ count}$ and $Train\ count = Trains\ run + Cancellations$, PPM worsens as cancellations increase. Punctuality is therefore a measure of trains which have passed PPM as a proportion of trains which have actually run on the network.

The main parameters of interest are the parameters affecting the 'slope' of the regression relationship (β_i^1, β_i^2 and β_i^3 in equation (1) and the analogous parameters in equations (2) and (3)). These measures capture how a change in PPM/Cancellations gives rise to a change in AML/DML, and are used later to 'translate' the PPM/CaSL trajectories into the 'language' of Schedule 8.

It should be emphasised that the above relationships are **not 'causal'** – performance minutes and the regulatory performance measures (PPM and CaSL) are simply alternative measures of performance. A separate set of fundamental processes – performance of signalling equipment, rolling stock reliability, weather, capacity utilisation and so forth – influence both types of measure via separate causal links. Such processes have not been explicitly taken into account in the model. A particular issues is that we have concluded not to include traffic growth in the model as this is captured implicitly in the PPM and CaSL trajectories. Future work could be undertaken to disentangle the effects in the future, but this is beyond the scope of the work to set the Schedule 8 benchmarks.

Given that the approach is to provide a framework for modelling correlation rather than causation – albeit by allowing correlations to take fairly 'general' forms (e.g. linear and

nonlinear) – it should be noted that statistical inference (t-statistics, F-statistics and so forth), loses its usual 'meaning', as emphasised see Steer Davies Gleave report.

A number of things should be noted about these equations:

i) The regression relationships contain, as default, a 'quadratic' term. This allows us the flexibility to estimate 'nonlinear' as well as 'linear' relationships.
ii) The equations include 'dummy' or 'binary' variables, which allow us to select a given period and test whether the relationship has changed following that period (for example as a result of a major timetable change or the introduction of new rolling stock). In particular, the equation includes both an 'intercept dummy' – which allows the overall level of the relationship to change following a given period – and a 'slope dummy' – which permits the slope or 'gradient' of the relationship between AML/DML and PPM/Cancellations to change following that period.
iii) Whilst it is not evident from the equations above, the Excel model has been constructed so that any given observation or set of observations can be omitted from the analysis, for example if such a period represents a genuine 'outlier' and appears to have an undue impact on the regression results.

We have sought to design the models to allow significant flexibility in undertaking the regressions. However, it is very important that the approach to making changes is objective and consistent, and that this flexibility is not misused. As such, we have followed the procedure described in the coming paragraphs to select the appropriate models.

As 'default', we have:

- Assumed a linear relationship (effectively omitting the quadratic term in the equations above);
- Used all data in the first four years of CP4; and
- Assumed that the relationship has been the same over the course of CP4 to date (i.e. assumed no structural changes).

In a number of cases, it has been necessary to deviate from this 'default' position in order to ensure that the relationships used to underpin the benchmarks are as robust as they can be. We have deviated from this standard approach **only** when the following 'dual' criteria are **both** met:

i) There must be a sound **operational rationale** for not following the 'default' approach (for example, using a nonlinear relationship should be supported by a clear operational justification); and
ii) There must be **robust statistical grounds** for not pursuing the 'default' approach. Although, as noted above, t-statistics lose their full interpretation in this context, since we are estimating correlation rather than causation, we still feel that they can be informative and impartial in deciding, for example, if structural breaks have occurred.

Prior to this written consultation, we have conducted a series of meetings and follow-up conversations with TOCs and Network Rail routes. This has allowed us to gain a better

understanding, and improve the robustness of the underlying regression relationships. Specifically, TOCs have provided a number of useful suggestions around the above adjustments, and where appropriate, we have reflected these suggestions in the analysis. We have set out the assumptions that we have made pertaining to each TOC in Appendix 1, and invite TOCs to check these assumptions and feed back any concerns they may have.

'Levels' versus 'differences'

The relationships are derived on the basis of a 'time-series' of data. That is to say, we derive the relationships on the basis of repeated observations of AML (and/or DML) and punctuality (and/or Cancellations) over time (this is in contrast to 'cross-sectional' analysis, in which we examine multiple units of study in the same period).

An important consideration is whether the relationships are derived on the basis of 'levels' (i.e. where we regress, say, DML on Cancellations) or 'differences' (e.g. where we regress the *change* in DML on the *change* in Cancellations).

During the project, we considered – and undertook some analysis – to inform whether the estimation should take place in levels or differences. We concluded that it is most appropriate to undertake the analysis in levels. The reasons for this are manifold:

- The levels approach is **simple**, **easily understood** and has received industry buy-in through the informal consultation process undertaken in advance of the formal consultation published in August.
- We note that the dispersion of the estimated slope parameters across Service Groups is significantly greater for the differences approach compared to the levels approach. In particular, the variance – the yardstick measure of statistical dispersion – is almost a third higher in the differences approach compared to the levels method (105 compared to 80 in the analysis we have undertaken). Similarly, the range of the betas is greater under the differences approach. The *minimum* beta is 2 in the differences approach, compared to 4 in the levels approach. And the *maximum* beta value is 68 in the differences approach, compared to 66 in the levels approach. We believe that these features offer a number of further reasons that suggest that the levels approach is likely to be more robust than the differences technique:
    - Whilst a degree of variation between the slope parameters across Service Groups is reasonable (reflecting local features), *a priori* one would expect that the dispersion should not be too great. As such, that the parameters are more consistent across Service Groups in the levels approach lends credibility to that method.
    - The lower dispersion in the levels approach means that the slopes will typically be 'closer' to the mean. To the extent that there may be anomalies in individual estimated slope parameters, using figures that are closer to the average is likely to reduce the risk of errors arising in setting trajectories. We believe that this will reduce financial risk to both operators and Network Rail.
    - The greater dispersion associated with the differences approach is likely to lead to less equitable outcomes across TOCs and Service Groups. In

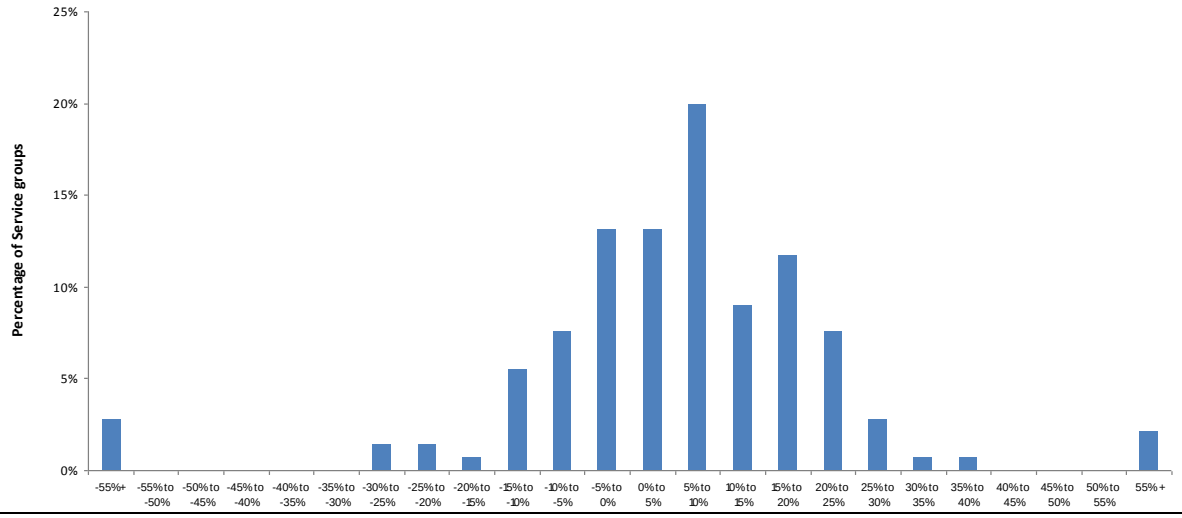
particular, under the differences approach, there will be a greater tendency for some TOCs and/or Service Groups' Schedule 8 benchmarks to respond strongly to regulatory trajectories, whilst other TOCs and/or Service Groups' benchmarks will respond weakly to those trajectories.

- The results of the regressions in levels and differences are **similar**. The bar chart, below (Figure 1), shows the results of estimating equation (3)[1] for all service groups (we have not labelled service groups for reasons of confidentiality). In approximately 81% of cases, the results for the slope coefficient – the key output of the analysis – for the first difference model are within 20% of the value for the level model.
- Differencing means a **loss of information**. By examining the period-by-period *change*, information about the *level* of the relationship is lost. Moreover, the act of calculating the difference means that the **sample size falls by one** (or more) in cases where data is not available for the entirety of the control period. This is a significant problem and this happens for Service Groups of a number of TOCs including Southeastern, Chiltern, Northern, ATW, Grand Central, First Transpenine Express, LOROL, and Virgin. Overall, approximately 10% of Service Groups are affected by this issue.
- The variables under consideration are **'cointegrated'** (see Technical Box), meaning that PPM and performance minutes appear to 'trend together' (and not 'drift apart') over time.
- There is a degree of **'arbitrariness'** about differencing. For example, the 'first difference' – whereby we examine how performance changed relative to the last period – is just one possibility. Given the strong seasonality in performance data, a more appropriate approach might be to use the 'thirteenth difference' – the change in performance since *the same period one year ago*. Similarly, taking the 'double-difference' or the 'difference in difference' also provides a consistent estimate of the slope parameter. So too does the 'triple-difference', the 'quadruple difference' and so forth. However, running the regressions on first difference, thirteenth difference or 'double difference' will not, in general, yield the same results.

We are not aware of any positive reasons to undertake the regressions in differences.

[1] We have used the full dataset (i.e. all periods in CP4 to date) with no structural breaks and have assumed a linear functional form for the purposes of this test.

Figure 1 – Percentage deviation of β from equation in differences relative to β from equation in levels



Technical Box 1 – Cointegration

In undertaking the regression in levels, it is clearly desirable that the series do not 'drift apart' over time. In order to test for this, we can appeal to the idea of 'cointegration'.

Formally, two variables are said to be cointegrated if their 'linear combination' is 'stationary'. Intuitively, this means that the variables do not tend to 'drift apart' over time, or equivalently that the error terms from the regression do not display any systematic trends.

We note that the cointegration test is simply a test of stationarity on the residuals of the regression output. As such, as a test of whether there is a 'drift' in the errors – and therefore a test of whether the variables drift apart over time – we believe that this is valid.

To test for cointegration, we can use the 'Engel-Granger' test (which is a special case of the 'Augmented Dickey-Fuller' or ADF test). If we denote the residuals from our performance minutes-PPM regression as e_i, the Engel-Granger test is conducted by running the following regression:

$$\Delta e_t = \theta e_{t-1} + \mu_1 \Delta e_{t-1} + \cdots + \mu_{t-k} \Delta e_{t-k} + \epsilon_t$$

The null hypothesis $H_0: \theta = 0$ is that the variables are **not cointegrated**. The alternative hypothesis $H_1: \theta < 1$ is that the variables **are cointegrated.** That is, in order to demonstrate that the variables are cointegrated – and therefore that it is appropriate to conduct the regression relationships described above without introducing problems – we must be able to reject the null hypothesis of no cointegration.

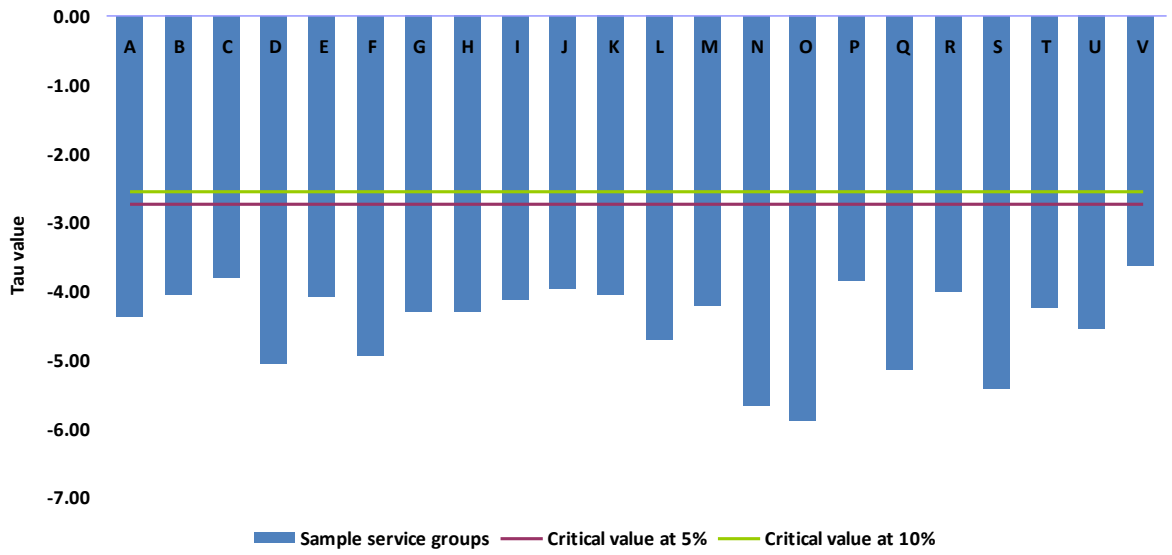
The graph, below, shows the results of the test for a sample of 22 service groups across all TOCs (we have used the ADF test with one lag, which appears to adequately remove all autocorrelation in the errors).

The 5% critical value is -2.76. The test statistics must be less than or equal to the critical

value to reject the null hypothesis that the variables are not cointegrated. For all 22 service groups we can reject the null hypothesis that the variables are not cointegrated[2]. For reasons of confidentiality, we have not labelled the service groups.

Overall, we believe that this evidence suggests that the variables are cointegrated, and this gives us confidence that the levels regressions are robust.

Figure 2 – Summary of Engel-Granger test results



Dependent and independent variables

Given that the relationships being estimated are around **correlation** rather than **causation**, at first sight it may not be obvious as to whether one should undertake the regressions with performance minutes on the 'left hand side' (LHS) (as in equations (1) to (3) above), or the 'right hand side' (RHS).

As it turns out, however, there is an unambiguous case for using performance minutes on the LHS. The reason is straightforward. Recall that we are using PPM (or punctuality or Cancellations, as appropriate) as the basis to 'predict' the measure of performance minutes. When applied with performance minutes on the LHS, OLS by definition, **minimises the errors made in generating a prediction of performance minutes on the basis of 'knowing' the regulatory measure (such as PPM and CaSL).** If, on the other hand, the regulatory measure was used on the LHS, OLS would minimise the errors made in generating a prediction of the regulatory measure on the basis of performance minutes – but the regulatory measure is already known (from the regulatory target) and performance minutes are unknown. A full illustration of this point is set out in Technical Box 2.

[2] The critical values are taken from J.Hamilton (1994), Time Series Analysis, Princeton University Press, P.766.

Technical Box 2 – Dependent and independent variables

The purpose of the exercise is to establish a prediction of performance minutes (or the change thereof) on the basis of the regulatory targets (or their change). The regulatory targets are, of course, defined in terms of PPM and CaSL (and by implication punctuality and cancellations). Therefore, in constructing our regressions, our aim is to minimise the errors made in predicting performance minutes on the basis of the regulatory measures.

To illustrate the issue, we use the notation y to represent our measure of Schedule 8 performance minutes (which is being 'predicted'), and x to represent our measure of the regulatory target (which is being used as the basis for the prediction). As such, our objective is to 'avoid' making prediction errors,

$$y - \hat{y}$$

where y is the actual value of the measure of performance minutes and $\hat{y}$ is the predicted value of y. The most common criterion used in statistics, which provides a measure of the 'average' error made in prediction is the 'Route Mean Squared Error' or RMSE,

$$RMSE = \sqrt{\frac{1}{n}\sum(y - \hat{y})^2}$$

Our objective, therefore, is to minimise the RMSE. We now examine two approaches – one in which y is regressed upon x ,and one in which x is regressed upon y – to illustrate the issue. We show that the first approach (when y is regressed on x) **always** yields a lower (or strictly speaking, 'no higher') RMSE than the alternative. We then provide an illustration of this result on the basis of the performance data.

Approach 1 – Regress y on x

Under this approach, we estimate the equation of the form

$$y = \alpha + \beta x + e$$

OLS, by definition, chooses estimates for α and β so as to minimise the RMSE. Put another way, the OLS estimates can be defined as,

$$\hat{\alpha}, \hat{\beta} = \text{argmin}\left(\sqrt{\frac{1}{n}\sum(y - \hat{y}_1)^2}\right) = \text{argmin}(RMSE)$$

where $\hat{y}_1$ is the predicted value of y on the basis of the above regression of Approach 1 (i.e. by using $\hat{\alpha}$ and $\hat{\beta}$). It follows that the resulting RMSE, call it $RMSE_1$, is equal to the *minimum possible R*MSE,

$$RMSE_1 = RMSE_{min}$$

Approach 2 – Regress x on y

Under this approach, we estimate the equation of the form

$$x = \gamma + \phi y + e$$

Applying OLS to this equation gives rise to estimates $\hat{\gamma}$ and $\hat{\phi}$. Then, rearranging the equation, the implied estimate of y on the basis of the known x is

$$\hat{y}_2 = \frac{\hat{\gamma}}{\hat{\phi}} + \frac{1}{\hat{\phi}}x$$

Note that, in general, $\hat{y}_2 \neq \hat{y}_1$. It follows that $RMSE_2 \neq RMSE_1$ and since $RMSE_1 = RMSE_{min}$, it must be the case that,

$$RMSE_2 \geq RMSE_1 = RMSE_{min}$$

In conclusion, we can say that the RMSE is minimised by pursuing Approach 1, and that Approach 2 will yield an RMSE that is no better, but could be worse, than Approach 1.

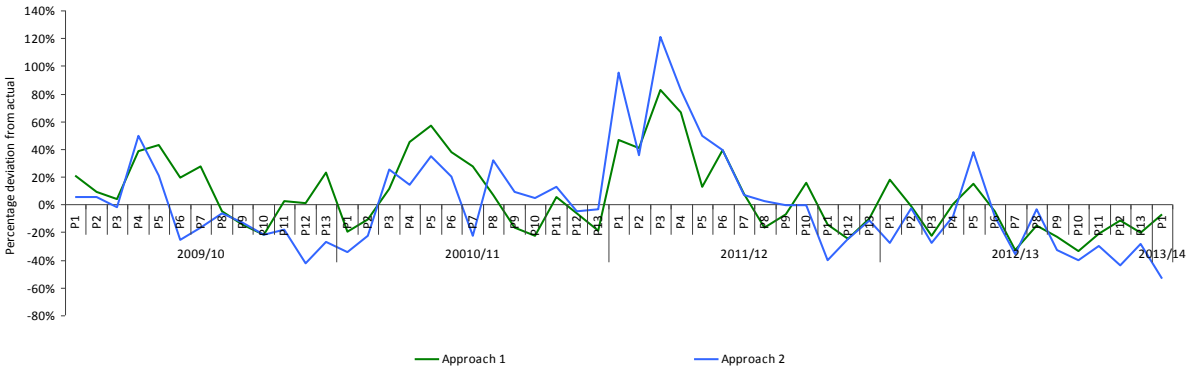
Illustration

To illustrate this point with some real figures, we have calculated the RMSE using Appraoch 1 and Approach 2 for all Service Groups, using total performance minutes and PPM. As expected, the RMSE was lower in **all cases** using Approach 1 compared to Approach 2. On average, the RMSE was around 250% greater using Approach 2 compared to using Approach 1. Even in the case when Approach 2 gave results that were closest to Approach 1, the RMSE was still 9% higher. In some cases, the RMSE from Approach 2 was almost 2000% higher than the RMSE from Approach 1.

Figure 3, below, provides an illustration for a single Service Group. It shows that Approach 1 predicts the actual performance minutes figure to a greater degree of accuracy than Approach 2.

We should be mindful, however, that the RMSE is just one measure of the potential error in the values predicted by a model and that, since it involves 'squaring' the errors, it can be sensitive to outlier data. A common alternative measure is the Route Mean Absolute Error (RMAE). As the name suggests, this measures the absolute error in the predicted values resulting from the model. We have compared the RMAE in both approaches, as described above. We also find that the RMAE is always lower in Approach 1 than in Approach 2.

Figure 3 – Deviation of prediction from actual performance minutes under both approaches



Deriving Schedule 8 benchmarks

We can use equations (1) to (3) to derive the change in the performance minutes as a result of a change in the regulatory variable. We illustrate this with respect to equation (1), but precisely the same arguments can be applied in respect of equations (2) and (3). Differentiating (1) with respect to $Punc_{it}$ tells us how AML changes as a result of a 'small change' in punctuality,

$$\frac{\partial AML_{it}}{\partial Punc_{it}} = \beta_i^1 + 2\beta_i^2 + \beta_i^3 D_{it} \qquad (5)$$

Of course, we do not know the 'true' values of the β parameters – these are the quantities that we have sought to estimate by means of regression analysis. However, our regression analysis gives rise to *estimates* of these quantities. Hence, the *estimated* change in AML_{it} as a result of a change in $Punc_{it}$ is,

$$\widehat{\frac{\partial AML_{it}}{\partial Punc_{it}}} = \hat{\beta}_i^1 + 2\hat{\beta}_i^2 + \hat{\beta}_i^3 D_{it} \qquad (6)$$

This means that, to a 'first order' approximation, a change in $Punc_{it}$ of $\Delta Punc_{it}$ leads to a change in AML_{it} of $\left(\hat{\beta}_i^1 + 2\hat{\beta}_i^2 + \hat{\beta}_i^3 D_{it}\right)\Delta Punc_{it}$.

Since the above equations tell us how AML/DML changes when PPM/Cancellations changes, we can use the relationships to derive the appropriate Schedule 8 benchmarks that 'mirror' the PPM/CaSL performance trajectories. In effect, the relationship has to be 'applied' twice:

i) once to adjust the Halcrow-derived AML benchmark from the 'baseline' period to the end of CP4 (CP4-exit); and
ii) again to adjust the CP4-exit benchmark derived in (i) to reflect the PPM/CaSL trajectories within CP5 itself.

Baseline period to CP4-exit

Halcrow has calculated baseline performance (expressed in terms of Schedule 8 AML and DML) as the average of performance in the 'benchmark period' of the two financial years 2010-11 and 2011-12. This baseline period was decided by ORR in consultation with the industry. For a small number of TOCs, a different baseline calibration period has been used, for example where a fundamental timetable change took place during the calibration period.

The first task is to use the statistical relationships derived above to establish the appropriate CP4-exit position (i.e. the year 2013-14). This is done using the following equation:

$$b_{iCP4-exit} = \overbrace{\widetilde{b_{i0}}}^{Baseline} + \left\{ \overbrace{(\hat{\beta}_i^1 + 2\hat{\beta}_i^2 Punc_{i0} + \hat{\beta}_i^3 D_{it})(\langle 1 - Punc_{iCP4-exit} \rangle - \langle 1 - Punc_{i0} \rangle)}^{Adjustment\ for\ change\ in\ Punc} + \underbrace{(\hat{\gamma}_i^1 + 2\hat{\gamma}_i^2 Canc_{i0} + \hat{\gamma}_i^3 D_{it})(CaSL_{iCP4-exit} - CaSL_{i0})\frac{Canc}{CaSL}}_{Adjustment\ for\ change\ in\ CaSL} \right\} \times NR\%$$

where $b_{iCP4-exit}$ is the CP4-exit performance minutes benchmark for service group i; b_{i0} is the Halcrow-derived baseline for service group i; $PPM_{iCP4-exit}$ is 2013-14 PPM for service group i; PPM_{i0} is PPM for that service group in the baseline period; and $NR\ \%$ is the 'percentage contribution' is the share of Network Rail total performance minutes in the benchmark calibration period (201-11 and 2011-12 typically).

Although it looks daunting, this equation simply says that the CP4-exit benchmark[3] is equal to the Halcrow-derived benchmark (from the calibration period 2010-11 and 2011-12) plus:

i) An adjustment to take account of changes in punctuality between the calibration period and the end of CP4; and
ii) An adjustment to take account of changes in Cancellations between the calibration period and the end of CP4[4].

The ratio $\frac{Canc}{CaSL}$ appears in order to 'remove' the Significant Lateness component from CaSL in order to avoid a 'double-count' – Significant Lateness is already captured as part of punctuality.

CP5 trajectory

Having established the CP4-exit level of AML/DML, we can move on to calculate the benchmarks for all years of CP5. To do this, we use the following equation:

$$b_{it} = \overbrace{\widetilde{b_{it-1}}}^{Previous\ year's\ benchmark} + \left\{ \overbrace{(\hat{\beta}_i^1 + 2\hat{\beta}_i^2 Punc_{i0} + \hat{\beta}_i^3 D_{it})(\langle 1 - Punc_{it} \rangle - \langle 1 - Punc_{it-1} \rangle)}^{Adjustment\ for\ change\ in\ Punc} + \underbrace{(\hat{\gamma}_i^1 + 2\hat{\gamma}_i^2 Canc_{i0} + \hat{\gamma}_i^3 D_{it})(CaSL_{it} - CaSL_{it-1})\frac{Canc}{CaSL}}_{Adjustment\ for\ change\ in\ CaSL} \right\} \times NR\%$$

This formula is identical in structure to the one above. It says that the benchmark in year t for the service group is equal to the benchmark in the previous year (i.e. in $t-1$), plus an

[3] Note that it is not possible to apply a *direct* adjustment to take account of *actual* patterns in AML/DML, since CP4-exit AML/DML is not forecast by the industry.
[4] For a small number of TOCs, only a PPM trajectory has been developed. In this case, only equation (3) is used to 'drive' the benchmarks.

adjustment for the change in punctuality since last year and an adjustment for the change in Cancellations since last year. Again, there is an adjustment for the Network Rail year-on-year contribution to the change in performance. The Network Rail contribution has been developed 'bottom-up' by Network Rail's routes, in consultation with TOCs. The assumptions used are set out in Appendix 1. We assume that 'joint' initiatives are split according to the percentage of performance minutes in the baseline period.

Since we only have a PPM and CaSL trajectory, it is necessary to approximate a Cancellations trajectory, and as such a punctuality trajectory. We have estimated this using the following equations:

$$Cancellations = CaSL * \frac{Cancellations_{baseline}}{CaSL_{baseline}}$$

$$Punctuality = PPM + Cancellations$$

Where the ratio of Cancellations : CaSL is calculated using data from the recalibration period.

Worked example of the benchmark calculation

To calculate the benchmark (i.e. the starting point that are consistent with JPIP) for CP4-exit, i.e. the year 2013-14, we use the following statistical relationship:

$$b_{iCP4-exit} = \overset{Baseline}{\widehat{b_{i0}}} + \left\{ \underbrace{(\hat{\beta}_i^1 + 2\hat{\beta}_i^2 Punc_{i0} + \hat{\beta}_i^3 D_{it})(\langle 1 - Punc_{iCP4-exit} \rangle - \langle 1 - Punc_{i0} \rangle)}_{Adjustment\ for\ change\ in\ Punc} + \underbrace{(\hat{\gamma}_i^1 + 2\hat{\gamma}_i^2 Canc_{i0} + \hat{\gamma}_i^3 D_{it})(CaSL_{iCP4-exit} - CaSL_{i0})\frac{Canc}{CaSL}}_{Adjustment\ for\ change\ in\ CaSL} \right\} \times NR\%$$

For this illustrative example, we assume the following parameters for service group i:

b_{io} = 5 (Halcrow baseline)

$\widehat{\beta}_i^1$ = 8 (output from regression analysis)

$\widehat{\beta}_i^2$ = 0 (output from regression analysis)

$\widehat{\beta}_i^3$ = 1 (output from regression analysis)

PPM_{i0} = 90% (PPM in the baseline period)

$PPM_{iCP4-exit}$ = 92% (PPM in the year 2013-14, i.e. CP4-exit rate)

$\widehat{\gamma}_i^1$ = 10 (output from regression analysis)

$\widehat{\gamma}_i^2$ = 0 (output from regression analysis)

$\widehat{\gamma}_i^3$ = 2 (output from regression analysis)

$CaSL_{iCP4-exit}$ = 2.8% (CaSL in the year 2013-14, i.e. CP4-exit rate)

$\mathrm{CaSL_{i0}}$ = 3.0% (CaSL in the baseline period)

$\mathrm{NR\ \%\ contribution}$ = 70% (share of Network Rail contribution to the change in performance)

$\frac{Canc}{CaSL}$ = Cancellation to CaSL ratio = 80%

Substituting in to the equation, we can calculate the baseline period to CP4-exit:

$$b_{iCP4-exit} = \overbrace{5}^{\text{Halcrow baseline}} + \left\{ \overbrace{((8+(2)(0)(0.9)+(1)(1)))(\langle 1-0.92\rangle - \langle 1-0.90\rangle)}^{\text{Adjustment for change in PPM}} + \underbrace{(10+(2)(0)(0.03)+(2)(1))(0.028-0.03)}_{\text{Adjustment for change in CaSL}}(0.8) \right\} \times 0.7$$

$$b_{iCP4-exit} = 5 + (-0.180 - 0.0192)\ \mathrm{x}\ 0.7$$

$$b_{iCP4-exit} = 4.8606$$

Therefore, using the Halcrow-derived baseline performance, and taking to account the changes in PPM and Cancellations between the calibration period and to the end of CP4 (as shown above), we arrive at a CP4-exit performance minutes (i.e. in the year 2013-14) of 4.8606.

Using the CP4-exit baseline of 4.8606, we calculate the benchmarks for the succeeding years of CP5, using the following equation:

$$b_{it} = \overbrace{b_{it-1}}^{Previous\ year's\ benchmark} + \left\{ \overbrace{(\hat{\beta}_i^1 + 2\hat{\beta}_i^2 Punc_{i0} + \hat{\beta}_i^3 D_{it})(\langle 1-Punc_{it}\rangle - \langle 1-Punc_{it-1}\rangle)}^{Adjustment\ for\ change\ in\ Punc} + \underbrace{(\hat{\gamma}_i^1 + 2\hat{\gamma}_i^2 Canc_{i0} + \hat{\gamma}_i^3 D_{it})(CaSL_{it}-CaSL_{it-1})\frac{Canc}{CaSL}}_{Adjustment\ for\ change\ in\ CaSL} \right\} \times NR\%$$

For this illustrative example, we will assume that Cancellations are constant, PPM improves by 0.5% every year and the Network Rail share of performance improvement is constant at 80% for each year, until the end of CP5.

Therefore, the baseline CP5 trajectory for each year is calculated as follows:

$$b_{2014-15} = \overbrace{4.8606}^{\text{Previous year's benchmark}} + \left\{ \overbrace{(((8+(2)(0)(0.9)+1))(-0.05)}^{\text{Adjustment for change in PPM}} + \underbrace{(((10+(2)(0)(0.03)+(2)(1))(0))}_{\text{Adjustment for change in CaSL}}0.8 \right\} \times 0.80$$

$$b_{2014-15} = 4.8606 + (-0.45+0) * 0.80$$

$$b_{2014-15} = 4.5006$$

The, above, equation shows that the baseline for the year 2014-15 is 4.5006.

$$b_{2015-16} = \overbrace{4.5006}^{\text{Previous year's benchmark}} + \left\{ \overbrace{(((8+(2)(0)(0.9)+1))(-0.05)}^{\text{Adjustment for change in PPM}} + \underbrace{(((10+(2)(0)(0.03)+(2)(1))(0))}_{\text{Adjustment for change in CaSL}}0.8 \right\} \times 0.80$$

$$b_{2015-16} = 4.5006 + (-0.45+0) * 0.80$$

$$b_{2015-16} = 4.1406$$

The, above, equation shows that the baseline for the year 2015-16 is 4.1406.

$$b_{2016-17} = \overbrace{4.1406}^{\text{Previous year's benchmark}} + \left\{ \overbrace{\big(((8+(2)(0)(0.9)+1)\big)(-0.05)}^{\text{Adjustment for change in PPM}} + \underbrace{(((10+(2)(0)(0.03)+(2)(1))(0))}_{\text{Adjustment for change in CaSL}} 0.8 \right\} \times 0.80$$

$$b_{2016-17} = 4.1406 + (-0.45 + 0) * 0.80$$

$$b_{2016-17=} 3.7806$$

The, above, equation shows that the baseline for the year 2016-17 is 3.7806.

$$b_{2017-18} = \overbrace{3.7806}^{\text{Previous year's benchmark}} + \left\{ \overbrace{\big(((8+(2)(0)(0.9)+1)\big)(-0.05)}^{\text{Adjustment for change in PPM}} + \underbrace{(((10+(2)(0)(0.03)+(2)(1))(0))}_{\text{Adjustment for change in CaSL}} 0.80 \right\} \times 0.80$$

$$b_{2017-18} = 3.7806 + (-0.45 + 0) * 0.80$$

$$b_{2017-18} = 3.4206$$

The, above, equation shows that the baseline for the year 2017-18 is 3.4206.

$$b_{2018-19} = \overbrace{3.4206}^{\text{Previous year's benchmark}} + \left\{ \overbrace{\big(((8+(2)(0)(0.9)+1)\big)(-0.05)}^{\text{Adjustment for change in PPM}} + \underbrace{(((10+(2)(0.03)(0)+(2)(1))(0))}_{\text{Adjustment for change in CaSL}} 0.80 \right\} \times 0.80$$

$$b_{2018-19} = 3.4206 + (-0.45 + 0) * 0.80$$

$$b_{2018-19} = 3.0606$$

The, above, calculations illustrate how we calculate the annual benchmarks. We note that, in this illustrative example, the calculations show that the benchmark in the last year of CP5 (i.e. the year 2018-19) is 3.0606.

Appendix 3

Steer Davies Gleave report

1 Introduction and Summary

1.1 Network Rail appointed Steer Davies Gleave to undertaken a review of the Network Rail Schedule 8 benchmarks spreadsheet model, as part of a review of income and Schedule 8 benchmarks models. The review is designed to establish that the model serves its intended purpose, that the theory behind it is sound, and that there are no material errors in calculations performed within the model. While the checks in the review have been thorough, the model does not have a detailed specification against which a comprehensive model audit could be undertaken, and input numbers have not been checked to their original source.

1.2 As agreed with Network Rail, we have performed a number of specific checks.

1.3 A separate Schedule 8 benchmarks model exists for each train operating company (TOC). The checks were initially performed on the Arriva Trains Wales model, then, at the suggestion of the Office of Rail Regulation, continued on the Greater Anglia model, which was chosen at random (but subject to having a reasonable number of different service groups on which tests could be performed) by SDG. Unless stated otherwise, the comments in this note pertain to the Greater Anglia model supplied by Network Rail on 8th August 2013.

1.4 Checks have been made to the models for the other TOCs. These checks took the form of "spot checks" which included all significant issues identified in the Arriva Trains Wales and Greater Anglia models, and included all TOC models. It the time available it was possible to check approximately 60% of issue x TOC combinations.

1.5 We have established that the model does serve its intended purpose as articulated by Network Rail.

1.6 We are satisfied that the theory behind the model and the assumptions within it are valid.

1.7 We are also satisfied that the current version of the Greater Anglia model is functioning properly, with no significant errors in the formulae used.

1.8 This note contains

- A description of the model purpose and structure
- A report on the top-down checks made
- A report on the bottom up checks made.

2 Model purpose and structure

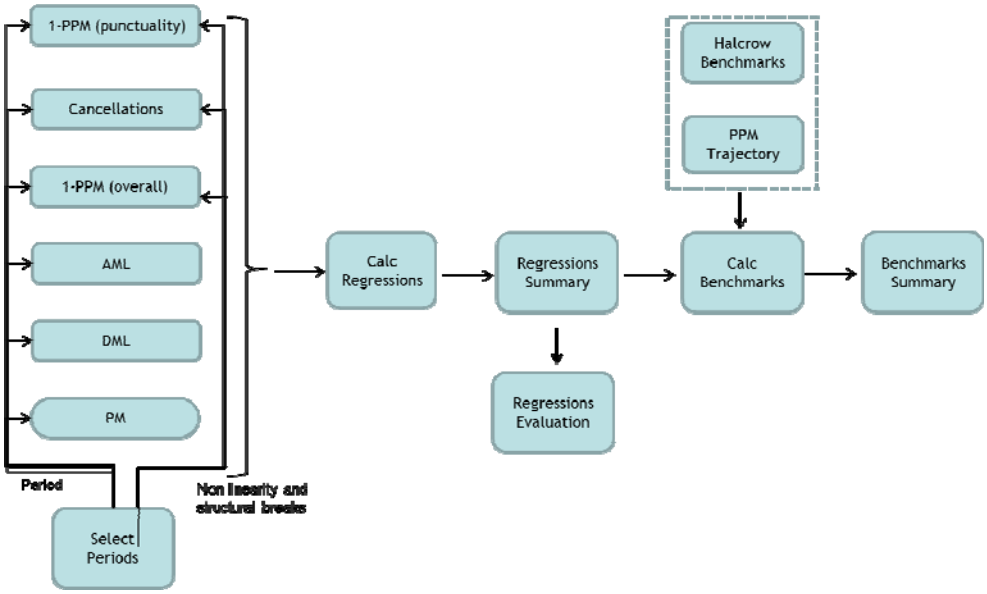
2.1 The purpose of the model is to establish Schedule 8 benchmarks for Network Rail for CP5 for each service group.

2.2 The model does this by:

- estimating the relationships between the Schedule 8 measure of performance i.e. 'average' minutes lateness (AML) and 'deemed' minutes lateness (DML) and measures of performance used for industry planning and regulatory targets i.e. Public Performance Measure (PPM) and Cancellations and serious lateness (CaSL); and then
- applying the relationships in order to 'translate' PPM and CaSL performance trajectories into Schedule 8 benchmarks defined in terms of the aggregate of AML and DML.

2.3 The theory behind the model and the methodology applied are set out by Network Rail in its "Setting Schedule 8 benchmarks for CP5 – Methodological note" which is contained in Appendix 2.

2.4 Figure 1 below shows a map of the model.



3 Top down checks

Does the model serve its stated purpose?

3.1 The model uses forecasts of PPM and cancellations to derive benchmarks. We note that in its letter to ORR of 14 June 2013 (Network Rail Schedule 8 benchmarks in CP5) Network Rail states that "Work is ongoing to model the relationships between PPM and performance minutes. These will be used to translate PPM trajectories into Schedule 8 benchmarks" and that ORR principles for setting Schedule 8 benchmarks (set out in its email of 14 August 2013 to industry representatives) include using relationships between Schedule 8 AML and the performance targets specified in the draft determination.

3.2 Given that the agreed industry approach is to use forecasts of PPM and cancellations to derive benchmarks, the model does serve its stated purpose.

Is the theory behind the model sound?

3.3 The theory behind the model is set out in the "Setting Schedule 8 benchmarks for CP5 – Methodological note". Essentially the model is based on a relationship between PPM and AML, and between cancellations and DML. It is a reasonable assumption that these variables will be correlated, however it is important to note that the relationship is not a causal one. Both PPM and AML are derivatives of delays, rather than AML being a derivative of PPM (or vice versa).

3.4 A number of specific questions on the theory behind the model are addressed below.

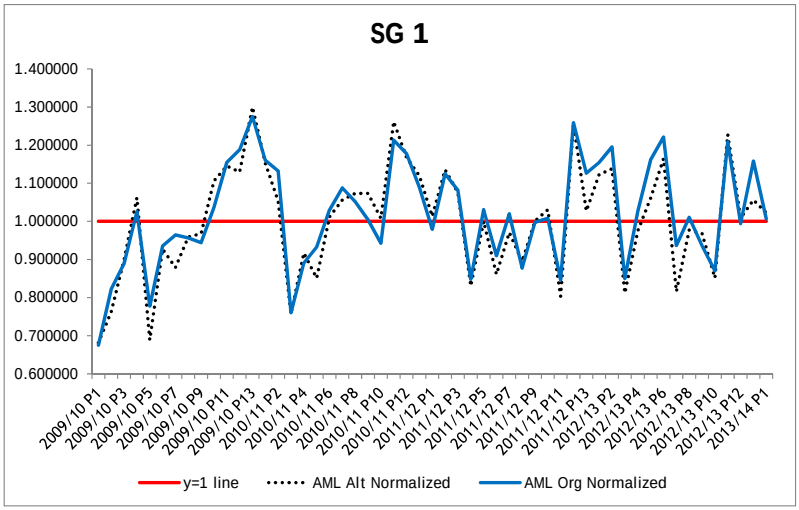
Which way round is the regression?

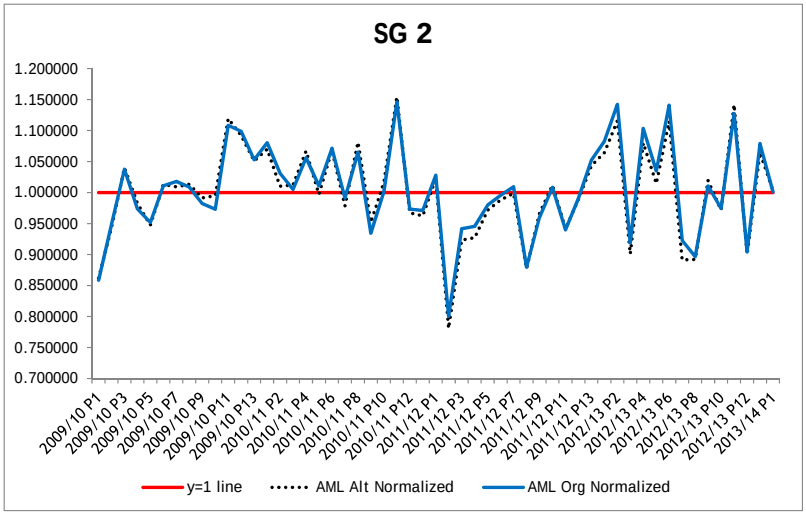
3.5 Given that the relationship between AML and PPM is not a causal one, it is not a foregone conclusion that the relationship should be established by regressing AML against PPM rather than vice versa. Network Rail justify their approach on the grounds that

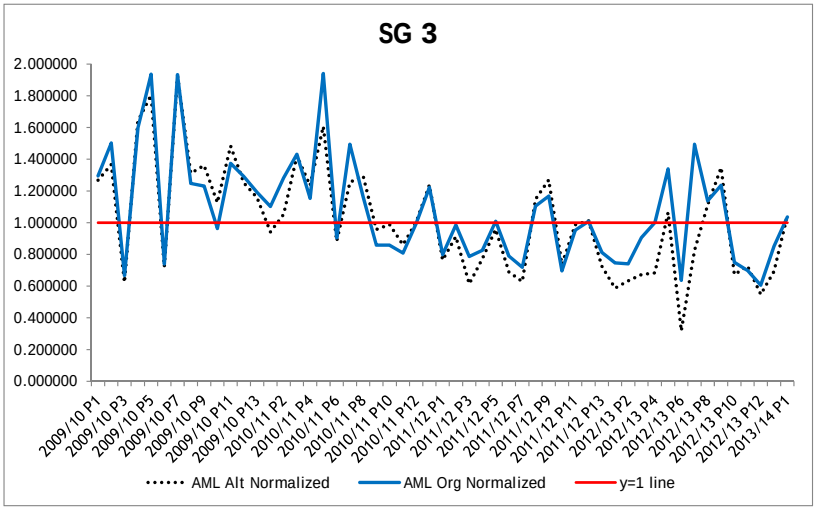
- the aim of the exercise is to predict the level of AML for a given PPM, so the method chosen is the more direct
- there is precedent in that the same approach was adopted in CP4.

3.6 We accept that these points are valid, and we have made a check by comparing the actual and predicted levels of AML for three random service groups (Figure 2). We have done this by predicting AMLs using the original (AML as the dependent variable) and alternative (PPM as the dependent variable) method. The graphs in Figure 2 normalise the predicted AMLs with respect to the actual AML (so we are comparing a series of Predicted AML/Actual AML). While it may not be immediately evident from the graphs, we have checked that the predicted AML using the regression with AMLs as the dependent variable results in lower mean square errors (MSE) than with the regression with PPM as the dependent variable. This is as expected.

Figure 2 TIME SERIES OF NORMALISED PREDICTED AML V ACTUAL AML USING ALTERNATIVE REGRESSION METHODS







Levels vs. differences.

3.7 The relationships are derived on the basis of 'levels' (i.e. where, for example, AML is regressed on PPM) as opposed to 'differences' (i.e. where, for example, the change in AML is regressed on the change in PPM).

3.8 In estimating the relationship between PPM and AML, Network Rail are not suggesting that they are estimating un underlying causal relationship. Rather, they have sought to define an intuitive relationship between two measures that can be expected to move together. Caution must be exercised in using statistical tests in these circumstances - the validity of the statistical tests depends on the properties of the error term, for example, whether there is a correlation between the error term and the regressors. However, we consider that Network Rail's use of the Augmented Dickey Fuller (ADF) test for cointegration (i.e. that the two variables are not diverging over time) is valid.

3.9 We are satisfied that using levels is a robust approach. As Network Rail point out, it has the merit (compared to using differences) of being more readily understood, and it also avoids the arbitrary choice of whether to take the first (i.e. one period) or thirteenth (i.e. one year) difference. Additionally, Network Rail's analysis shows that the variance in the coefficient on PPM is larger in a difference approach than a levels approach. This implies that there is more risk associated with errors arising from the setting of benchmarks if a difference rather than levels approach is used. We therefore support the use of this approach.

Are the assumptions in the model correct?

3.10 The assumptions in the model are reasonable. Two significant inconsistencies were found (these were identified in the review of the ATW model)

- there was an inconsistency between the regression used to derive the relationship and the way in which the relationship was applied
- the calculation of punctuality was inconsistent with its definition.

These inconsistencies have been removed, and the model is now internally consistent.

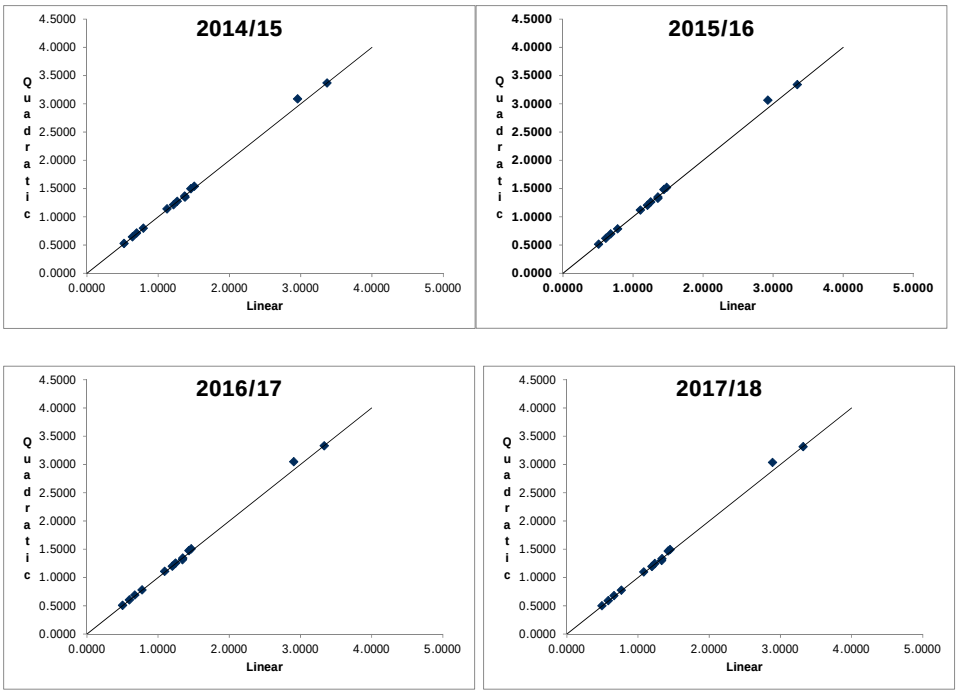
Is the relationship linear?

3.11 The model assumes that (within the ranges of PPM observed) the relationship between PPM and AML is linear. If in fact the relationship were non-linear and the graph of AML vs. PPM were either consistently concave or consistently convex, this could cause bias in the forecast benchmarks.

3.12 We have tested this by comparing the regressions with and without a quadratic PPM term.

3.13 We plotted scatter diagrams for benchmarks calculated using only the linear term and for benchmarks calculated using a quadratic term in addition to the linear term (both regressions have a constant), by service group. This is done for four of the CP5 years (Figure 2). We find that including the quadratic term does not systematically over or underestimate the benchmarks for service groups. There is a near perfect correlation coefficient of +1 between the two groups of regressions (linear and quadratic). We therefore conclude that it is not inappropriate to exclude the quadratic term from the model.

Figure 3 SCATTER DIAGRAMS OF LINEAR AND QUADRATIC BENCHMARKS – FOR A SAMPLE OF SERVICE GROUPS



Have the correct inputs been used and incorporated appropriately?

3.14 Source spreadsheets for all the inputs have been provided, and these have been correctly transferred to the model. The PPM and CaSL figures are sourced from TRUST, and the minutes lateness figures are sourced from PEARS. These are the appropriate sources.

Sense check by replicating calculations

3.15 We have sampled four of the 13 service groups in the Greater Anglia model (counting 'peak' and 'off peak' as separate service groups) regressions either by checking the formulas or constructing the regression in excel from scratch and found that the spreadsheet model is performing the regressions correctly.

4 Bottom up checks

Automated checks

4.1 We have undertaken a number of bottom up checks of the spreadsheet, assisted by the Spreadsheet Professional software, which checks for

- consistency of formulae across rows and columns within worksheets
- lack of circular references
- unused input values
- errors in range names or external links.

4.2 There are no errors in any category above in the current version of the model.

Manual checks

4.3 We have also made a number of manual checks

- are formulae correct and reproduced accurately?
- do the switches to exclude certain periods work correctly?
- are inputs transcribed correctly?

4.4 Our checks have shown that the formulae in the spreadsheet are consistent with the specification in the methodology statement, and thus the model is correctly carrying out the intended calculations.

4.5 The switches which exclude certain periods from the regression calculation (where data may be distorted due to exceptional circumstances) work correctly.

4.6 The inputs have been correctly transferred to the model.

Appendix 4

Draft determination consistent CP5 Schedule 8 benchmarks

TOC-specific