Technological University Dublin
ARROW@TU Dublin

2013

A Window of Opportunity: Assessing Behavioural Scoring

Kenneth Kennedy
Technological University Dublin, kennedykenneth@gmail.com

Brian Mac Namee
Technological University Dublin, brian.macnamee@tudublin.ie

Sarah Jane Delany
Technological University Dublin, sarahjane.delany@tudublin.ie

See next page for additional authors

Follow this and additional works at: https://arrow.tudublin.ie/scschcomart

Part of the Other Computer Engineering Commons

Recommended Citation

Kennedy, K., Mac Namee, B., Delany, S. J., O'Sullivan, M., & Watson, N. (2013). A window of opportunity: Assessing behavioural scoring. Expert Systems with Applications: An International Journal, 40(4), 1372-1380. doi:10.1016/j.eswa.2012.08.052

This Article is brought to you for free and open access by the School of Computer Sciences at ARROW@TU Dublin. It has been accepted for inclusion in Articles by an authorized administrator of ARROW@TU Dublin. For more information, please contact arrow.admin@tudublin.ie, aisling.coyne@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 4.0 License

Funder: ABBEST

Authors

Kenneth Kennedy, Brian Mac Namee, Sarah Jane Delany, Michael O'Sullivan, and Neil Watson

This article is available at ARROW@TU Dublin: https://arrow.tudublin.ie/scschcomart/18

Main
Click here to view linked References

A Window of Opportunity: Assessing Behavioural Scoring

K. Kennedy[a,*], B. Mac Namee[a], S.J. Delany[a], M. O'Sullivan[b], N. Watson[b]

[a] School of Computing, Dublin Institute of Technology, Ireland
[b] Irish Credit Bureau, Dublin, Ireland

Abstract

After credit has been granted, lenders use behavioural scoring to assess the likelihood of default occurring during some specific outcome period. This assessment is based on customers' repayment performance over a given fixed period. Often the outcome period and fixed performance period are arbitrarily selected, causing instability in making predictions. Behavioural scoring has failed to receive the same attention from researchers as application scoring. The bias for application scoring research can be attributed, in part, to the large volume of data required for behavioural scoring studies. Furthermore, the commercial sensitivities associated with such a large pool of customer data often prohibits the publication of work in this area. This paper focuses on behavioural scoring and evaluates the contrasting effects of altering the performance period and outcome period using 7-years worth of data from the Irish market. The results of this work indicate that a 12-month performance period yields an easier prediction task when compared with other historical payment periods of varying lengths. This article also quantifies differences in the classification performance of logistic regression arising from different outcome periods length. Our findings show that the performance of a logistic regression classifier degrades significantly when the outcome window is extended beyond 6-months. Finally we consider different approaches to how the concept of default is defined. Typically whether the customer is identified as a default risk or not is set based on either (i) whether the account is in default at the end of the outcome period or (ii) at any time during the outcome period. This paper studies both approaches and finds that the latter approach resulted in an easier classification problem, that is, it gives the highest assurance that the classification will be correct.

Keywords: Data Mining, Supervised Classification, Behavioural Scoring, Credit Scoring

1. Introduction

The term *credit scoring* is used to describe the process of evaluating the risk an applicant poses of defaulting on a financial obligation (Hand and Henley, 1997). The objective is to assign customers to one of two groups: *good* and *bad*. A member of the good group is considered likely to repay their financial obligation. A member of the bad group is considered likely to default on their financial obligation. In its simplest incarnation a credit risk scorecard consists of a set of characteristics that are used to assign a credit score to a customer indicating their risk level. This credit score can then be compared with a threshold in order to make a decision. As credit scoring is essentially a discrimination problem (good or bad), one may resort to the numerous classification techniques that have been suggested in the literature (see Lee *et al.*, 2005).

Based on both the task and data used, credit scoring is traditionally divided into two broad types (Bijak and Thomas, 2011). The first, *application scoring*, is used at the time an application for credit is made and estimates an applicant's likelihood of default in a given time period. The data used for model fitting for this task generally consists of financial and demographic information about a sample of previous applicants along with their creditworthiness at some later date. The second type of credit scoring, *behavioural scoring*, is used after credit has been granted and estimates an existing customer's likelihood of default in a given time period. Behavioural scoring allows lenders to regularly monitor customers and help coordinate customer-level decision making. The data used for model fitting for this task is based on the customers' loan repayment performance and also their creditworthiness at some later date. To be profitable a bank must accurately predict customers' likelihood of default over different time horizons (1 month, 3 months, 6

---

[*] *Correspondence: Kenneth Kennedy, K107A, School of Computing, Dublin Institute of Technology, Kevin St., D8, Ireland*
Email address: `kennedykenneth@gmail.com` (K. Kennedy)

months, etc.). Customers with a high risk of default can then be flagged allowing the bank to take appropriate action to protect or limit itself from losses.

The rest of this work focuses on credit risk-based behaviour scorecards, specifically account management scorecards which allow financial institutions to determine creditworthy and valuable customers in a timely and accurate manner. Behavioural scoring is used by organisations to guide lending decisions for customers in: credit limit management strategies; managing debt collection and recovery; retaining future profitable customers; predicting accounts likely to close or settle early; offering new financial products; offering new interest rates; managing dormant accounts; optimising telemarketing operations; and predicting fraudulent activity (Hand and Henley, 1997; McNab and Wynn, 2000; Sarlija *et al.*, 2008; Malik and Thomas, 2010; McDonald *et al.*, 2011).

To build behavioural scoring models practitioners must make decisions on a number of important parameters. This involves asking pertinent questions such as: The extensiveness of the historical data with which to model customer performance? How far forward into the future to make reliable predictions? What defines a loan defaulter? The credit scoring literature does not contain strong recommendations on how to answer these questions. This paper investigates some of the main issues affecting the construction of behavioural scoring models by examining the performance of retail loans issued by the main Irish banks in 2003 and 2004. The findings reported in this paper are based on real world data from a credit bureau. Credit bureaus are institutions that collect data on the performance of loans granted by different lenders. First, we compare the accuracy of scoring models that are built using different durations of historical customer repayment data (6-months, 12-months, and 18-months). Next, we quantify the differences between varying outcome periods from which a customer's default status (i.e. good or bad) is defined (3-months, 6-months, 12-months, 18-months, and 24-months). Finally, we demonstrate differences between alternative approaches used to define the customer's default status. A customer's class label of good or bad is normally defined by the number of missed loan repayments (i.e. the number of months or days in arrears). Typically, two common approaches are used to classify a customer as bad are if they are in arrears at either the *end of* or *during* some specified period, both of which are evaluated in this paper.

The rest of this paper is structured as follows, in Section 2 we summarise the relevant research on behavioural scoring. Section 3 describes our experimental set-up, including the data used and the experimental methodology. In Section 4 we describe our experimental results. Finally, in Section 5 we reflect upon the implications of our findings for the Irish mortgage market and describe directions for future research.

2. Background

The first behavioural scoring system to predict credit risk of existing customers was developed by Fair Isaac Inc. for Wells Fargo in 1975[1]. Behavioural scorecards have since evolved to influence decisions across the entire credit cycle. For example, *usage scorecards* for credit card products attempt to predict future levels of activity to assist in retention and incentive strategies. *Account management scorecards* are used during the lifetime of an account by lenders to predict the risk of default at a given point in time (e.g. every month, quarter, year). This allows the lender to set loan limits on top-up loan decisions and take appropriate measures to contain bad, loss-making accounts. Anecdotal evidence suggests that small business owners and start-ups frequently use personal consumer credit cards to pay for business expenses. As a result of such actions, lenders commonly use a mix of different customer behaviour models to assist lending decisions and set terms and conditions attached to the loan product (Finlay, 2010). Such information is also valuable to lenders' marketing departments when selecting profitable customers for additional products or calculating to what extent to incentivise increased account usage. A detailed list of the different types of behavioural scorecards is provided in McNab and Wynn (2000).

The two main types of statistical models used in credit behavioural scoring are duration models and classification models (Medema *et al.*, 2009). Discriminant analysis and logistic regression are both classification models and regarded as the standard industry approach for constructing behavioural scoring systems (Stepanova and Thomas, 2001; Andreeva, 2005; McDonald *et al.*, 2011). In duration models, the focus is not whether an applicant will default, but if they default when will this occur (Banasik *et al.*, 1999). Survival analysis is one such example of a duration modelling approach. In recent years, researchers have begun to investigate the applicability of techniques developed in survival analysis to predict the time to default (Andreeva *et al.*, 2005). The most frequently used model for survival analysis, the proportional hazards model (also called the Cox model), performs competitively with logistic regression in predicting customers' creditworthiness (Andreeva, 2005; Baesens *et al.*, 2005).

---

[1] `http://www.fico.com/en/Company/Pages/history.aspx`

Broadly speaking, behavioural scoring models can use static characteristics about the customer's past performance; or techniques which incorporate dynamic aspects. Thomas *et al.* (2001) survey the approaches and objectives of behavioural scoring, with particular focus on procedures that incorporate dynamic aspects of customer behaviour. The findings reported in this work are based on the use of static characteristics in behaviour scorecards. As we are using real world data from a credit bureau our work focuses on a special type of behavioural scoring called *credit bureau scoring* (see Bijak and Thomas, 2011), however our findings are applicable to the many types of behavioural scoring.

Figure 1 illustrates the longitudinal aspect to the data used in behavioural scoring. A sample of customers is selected so that the data on their product (e.g. loan) performance either side of an arbitrarily chosen *observation point* is available (Thomas *et al.*, 2001). The period before the observation point is often termed the *performance window*. Data on the customers' performance during this time is structured into features which are used by the behavioural scoring system to distinguish between customers' likely to repay their loan and those likely to default on their financial obligation.

Insert Figure 1 here

Figure 1: Behavioural scoring performance window and outcome window.

The data used in the performance window is derived from the banks' own internal databases and external data sources such as credit bureaus. The data includes customers' factual (e.g. date of birth, address), transactional (e.g. purchase history), and performance (e.g. arrears) characteristics. Based on the work of McNab and Wynn (2000), Table 1 lists the sources of typical behavioural scoring features. The appropriateness of the features will vary depending on the behavioural scoring system. For example, features from the *Promotions history* are less appropriate for behavioural scoring retail loans with fixed long term repayments in comparison to attrition scorecards that attempt to predict accounts that are likely to close, become inactive or settle early.

Table 1: Behavioural scoring data sources and associated features (McNab and Wynn, 2000).

Insert Table 1 here

The period after the observation point is known as the *outcome window*. The purpose of the outcome window is to classify borrowers into distinct populations (i.e. good and bad) based on their level of arrears (Mays, 2004). Selecting an appropriately sized outcome period requires careful consideration. As this period of time is used to classify customers, a comprehension of economic conditions, lender policies, and borrower volatility is necessary. If this period of time occurs during favourable economic conditions, then the performance of the scoring model may degrade if the reverse is true. Financial institutions need to consider the effects on customer behaviour caused by adopting certain operational policies, or *policy bias* Thomas (2009). For example, an early intervention policy may reduce the incidence of customers missing further loan repayments and subsequently being classed as bad. Finally, the outcome period should be sufficiently sized so as to capture a representative sample of bads with which to build a stable behavioural scoring model.

How a bad is defined depends on the objectives of the behavioural scoring system and the financial institution's view of success or failure (McNab and Wynn, 2000). According to Anderson (2007), in credit risk-based behaviour scorecards financial institutions can choose between: (i) a *current status* label definition approach that classifies a customer as either good or bad based on their account status at the end of the outcome window; and (ii) a *worst status* label definition approach which classifies a customer as either good or bad based on their account status during the outcome window. Commonly, as per Basel II (Basel Committee on Banking Supervision, 2006), a customer 90-days *worst status* is considered bad. The *current status* label definition is often used when managing early-stage delinquencies (Anderson, 2007).

Typically, the same techniques (e.g. logistic regression) are used by application scoring and behavioural scoring that best classify customers into one of two categories: goods and bads (Thomas *et al.*, 2001). Behavioural scorecard modellers encounter many of the same scorecard development and implementation issues as with application scoring, such as: identifying and adjusting for different segments of the population (see Bijak and Thomas, 2011), ensuring the optimal correlation between features (see Tsai, 2009), handling class imbalance (see Burez and den Poel, 2009), identifying the correct sample size (see Crone and Finlay, 2011), to name but a few challenges.

This study addresses issues affecting the performance of logistic regression as it remains the most widely used approach for behavioural scoring. There is no standard way to define the length of the performance and outcome windows. The recommendations in the literature typically range from 6-to-24-months (see Thomas *et al.*, 2001, 2002; Mays, 2004; Thomas, 2009; van Gestel and Baesens, 2009). Practitioners generally use a number of different candidates (e.g. 12, 18, and 24 months) until it is clear which one performs best. Practitioners also need to consider that using too long a performance window may result in customer dissatisfaction about the length of time it takes for their credit scores to improve (SAS, 2009). Clearly this is a time consuming task with no guarantee that the selected scorecard will fit the data adequately. For example, as previously mentioned, the outcome window should be long enough to capture most of the bad events. This can be attempted by plotting the cumulative default rate curve by time and measuring the percentage of bads captured as the outcome window size increases. Consequently, a trade-off needs to be found between an outcome window of reasonable length and a well performing model. Consider a 10-year outcome window which may well capture every type of customer volatility since the observation point, but the population that is then applied to the behavioural scorecard may be different from that which it is built on. Such undetected faults can have deleterious effects on scorecard reliability.

Using real world data this study evaluates the differences in model performance using different sized performance and outcome windows. The scope of the study is enhanced by comparing the window sizes using both *current status* and *worst status* label definitions.

Till and Hand (2003) observe that much interesting work involving datasets whose contents are confidential fails to reach publication. This results in wasted effort repeating the same or similar assessments used to measure models' effectiveness and accuracy, in addition to techniques used to attract and retain profitable customers (Till and Hand, 2003). To the best of our knowledge, to date very little empirical research has been published in the literature investigating the effects of different sized time horizons on classifier performance. Much of the recent research for determining appropriate time periods in behavioural scoring is conducted in the context of assessing the applicability of survival analysis as a method of identifying loan defaulters. This is often performed by comparing the performance of a survival analysis based technique with a classification approach, typically logistic regression.

The next section will explain the experimental design in more detail.

3. Experiment Set-up

The aims of this evaluation described are to evaluate the efficacy of an assorted set of performance and outcome window sizes on the classification accuracy of logistic regression. This is achieved by creating multiple behavioural scoring datasets with varying performance and outcome window sizes. This study also assesses two popular approaches used to label behavioural scoring data, namely the *current status* label definition approach and the *worst status* label definition approach. To accomplish these aims we adopt three overlapping experiments:

(i) A comparison of performance window sizes by classifying loans over a range of fixed outcome window sizes and varying performance window sizes.

(ii) A comparison of outcome window sizes by classifying loans using a single fixed performance window size and varying outcome window sizes.

(iii) A comparison of label definition approaches by classifying loans for both label definition approaches using a single fixed performance window size and varying outcome window sizes, as per (ii).

The first stage of the analysis, (i), considers the average class accuracy of the logistic regression model constructed using a selection of performance window sizes. The comparison is performed using the *worst status* label definition approach. Based on the results of (i) the best performing performance window size is then used in (ii). The outcome window sizes are compared using both the *worst status* approach and the *current status* approach. The final stage of the analysis, (iii), performs a direct comparison of the *current status* and *worst status* label definition approaches.

The comparison is performed using the datasets generated for (ii). In the study the performance window sizes used are 6-months (*P6*), 12-months (*P12*), and 18-months (*P18*). The five outcome window sizes used are 3-months (*O3*), 6-months (*O6*), 12-months (*O12*), 18-months (*O18*), and 24-months (*O24*). This section describes the data sets, performance measures and methodology used.

3.1. Data

We have used data provided by the Irish Credit Bureau (ICB). The data was anonymised to protect customer confidentiality and identity. It contains 2,500 customers who were approved between January 2003 and December 2004. The data provided includes a subset of their application characteristics and full subsequent repayment behaviour up to December 2010. Based on statistics issued by the Irish government's Department of Environment, Community and Local Government (DofE, 2008), the number of Irish mortgages issued in 2003 stood at 97,726 (or €17,432m worth) and for 2004 the figure recorded was 104,134 (or €21,003m worth). As such, this data represents a random sample of just over 1% of all mortgages issued in Ireland between 2003 and 2004. Each data record details a customer's repayment behaviour for the previous 24-months. Typically the ICB receive monthly data record updates; in this dataset each data record is updated every 3-months resulting in multiple data records per customer.

Table 2 describes the features of each ICB data record. The features are grouped into customer loan application data (*Application data*) and customer repayment behaviour data (*Behavioural data*). The application data for each data record remains unchanged from the time of the original loan application. The behavioural data is updated every 3-months, as specified by the *Account update date*. *Loan Protection* is used to indicate if the customer has some form of financial protection against an unforeseen adverse personal event. The *Account association* feature indicates if the loan is associated with a single or joint account. The *Outstanding loan balance* is the overall amount owed. Also included is a non-standard ICB feature, *Loan installment amount*, which is the amount repaid since the last quarter. This is calculated as the difference between the current *Outstanding loan balance* and that of the previous quarter.

Table 2: ICB data features. Features removed from the ICB data during data preparation are indicated by *.

Insert Table 2 here

Current credit bureau information is contained in the *repayment indicators* feature. These features record not only the current status of the loan, but also the loan status for each of the previous 23 months. We do not disclose the features due to commercial confidentiality. The performance window size dictates how many months worth of the repayment indicators are used to train and test the model.

In order to obtain a final dataset with which to train and test a logistic regression classifier it is necessary to clean and prepare the ICB dataset. We used the existing ICB features from Table 2 (except those with $*$) and generated additional features. An additional set of features are derived from the repayment indicators to form the *combination features*. The combination features describe the state of the account over the performance window with respect to a particular repayment indicator. In total, 61 combination features are defined. For example, Table 3 lists the combination features derived from the *arrears repayment indicator*.

Table 3: Combination features generated based on the arrears repayment indicator. This example is based on a 18-month performance window.

Insert Table 3 here

The arrears repayment indicator is used to create features to indicate the following: that the account has been in arrears sometime during the performance window; arrears have occurred over certain times frames (e.g. in the last 3-months); the highest number of missed payments; only one missed payment; a missed payment on two separate occasions; that the account has moved from arrears to current. The size of the performance window may preclude certain features from being used (e.g. arrears in months 13-to-18).

The instances in the generated dataset need a label (good or bad) in order to build a model. A customer is defined as bad if they are 60-days or more in arrears on their loan, i.e. two or more missed monthly payments. When a customer is defined as bad using 90-days or more in arrears on their loan (i.e. three or more missed monthly payments) too few defaulters are generated with which to train the logistic regression model. The *current status* approach *worst status* approach are used to define our class label. Experiment results are reported and compared for both approaches.

Insert Figure 2 here

Figure 2: Experiment set-up for a 12 month performance window and a 3-month outcome window (Out).

During the data cleaning stage features that uniquely identify the data record or customer were removed (see Table 2). Any data record with missing or incomplete information was removed. After data preparation and data cleansing we obtained a final usable dataset of close to 45,000 instances with each instance representing a quarterly snapshot of customer behaviour over the last 24-months. The next section describes how this dataset is used to create training and test datasets as specified by performance window and outcome window sizes.

3.2. *Methodology*

The training data, comprising solely of customer accounts opened in 2003, was divided into two subsets: (i) the classifier training set (67%); and (ii) the validation set (33%). The classifier training set and the validation set were used to train and tune the classifier. Test sets were used to evaluate classifier performance and were comprised solely of customer accounts opened in 2004. This ensured that the customer accounts in the training and test sets were mutually exclusive of each other. The performance window start date was set as the beginning of January 2005 for all of the initial test sets. For subsequent test sets the performance and outcome window start dates were incremented by 3-months, e.g. for the second test set the performance window start date moved to the beginning of April 2005. Classifier performance was measured and the process was repeated until the outcome window end date reached the end of 2010.

Figure 2 displays the experimental set-up for a 12-month performance window and 3-month outcome window. For the training data, the performance window start date was set as the beginning of January 2004. The performance window end date, determined by the performance window size, was set as the end of December 2004. The outcome window began immediately after the performance window end date and its end date, determined by the outcome window size, was set as the end of March 2004.

It was necessary to exclude a number of instances from the datasets used to train and test the model. If multiple instances of the same customer account occurred during the performance window, the most recent instance (i.e. closest to the performance window end date) was used. Any instances less than 3 months old (measured using the account opening date and the performance window end date) were also removed (Mays, 2004, pp.151). Instances marked as closed from the beginning of the performance window were also removed. In order to maximise the separation between the goods and the bads, indeterminates were omitted from the data used to train the model. An indeterminate is defined as someone whose behaviour is somewhere between good and bad, i.e. 1 month in arrears (Thomas *et al.*, 2001). Indeterminate accounts were kept in the test data.

A logistic regression classifier was trained and tested on our real world dataset. As already stated, logistic regression was selected as it is widely used to build credit scoring models (Hand and Zhou, 2009). Despite the existence of more sophisticated classification models such as support vector machines and neural networks, the popularity of logistic regression has endured. This may possibly be due to the interpretability and fast estimation of its parameters. The logistic regression model was implemented using the Weka (version 3.7.1) machine learning framework (Hall *et al.*, 2009). Tuning involved optimising the logistic regression ridge estimator parameter in order to offset unstable coefficient estimates that arise from highly correlated data. Each experiment was conducted 10 times using different randomly selected training and validation set splits and the results reported are averages of these 10 runs.

3.2.1. *Performance Measures*

In practice it is necessary to select a threshold on classification output in order to make actual classifications. Classifier performance was assessed using the average class accuracy which measures classification performance at a specific classification threshold. The validation data set was used to identify an optimised threshold.

Differences between the performance of the logistic regression classifier on various test datasets were analysed with a Kruskal-Wallis one-way analysis of variance by ranks and *post hoc* pairwise comparisons performed with a Dwass-Steel-Critchlow-Fligner procedure (see Wolfe and Hollander, 1999). Kruskal-Wallis one-way analysis of variance by ranks is a non-parametric method for testing whether k groups have been drawn from the same population. If the result of the Kruskal-Wallis one-way analysis of variance by ranks is significant, it indicates that there is a significant difference between at least two of the groups in the set of k groups (Sheskin, 1997). As the test is non-parametric, it makes no assumption about the shape of the population distribution from which the groups are drawn. When significant differences were indicated by Kruskal-Wallis, all possible pairwise comparisons between the groups were made using the Dwass-Steel-Chritchlow-Fligner *post hoc* test.

A two-tailed Mann-Whitney U test (see Wolfe and Hollander, 1999) was used to test for statistically significant differences between the average class accuracies of the *worst status* and *current status* label definition approaches. Statistical significance was established at $p < 0.05$ and statistical analysis was performed using the statistical package StatsDirect (Buchan, 2011).

Section 4 will describe the results of this experimental process.

4. Results and Discussion

The objectives of this evaluation were threefold. First we examined the most appropriate performance window size. Next we determined the differences between outcome window sizes. Lastly we compared two commonly used labelling approaches.

4.1. Performance Window Selection

Figure 3 provides a graphical representation of the results for each performance window over the five separate outcome window sizes using the *worst status* label definition approach. From Figure 3, a number of general points are apparent. First, the 12-month performance window outperformed both of the other performance windows (P6 and P18) when used with shorter outcome windows (i.e. O3, O6, and O12). This is interesting as it highlights how a 6-month performance window is effected by uncaptured seasonal factors. For example, a once-off marketing campaign aimed at attracting a specific customer profile may result in different characteristics from those of future expected applicants. It is also likely that the 6-month performance window is too short to allow for a sufficient accumulation of transitions in customer repayment behaviour with which to build a stable classification model. Conversely, the 18-month performance window maybe too long and events occurring earlier in the performance period should not be given the same weight or importance as more recent transitions. Secondly, frequently the average class accuracy of the 6-month performance window was unable to achieve parity with the longer performance windows. This relates to the previous point regarding the stability of the classification model.

Insert Figures 3a - 3e here

Figure 3: Average class accuracies (y-axis) of each outcome window. *Worst status* label definition.

For each outcome window in Figure 3 differences in classifier performance were tested with the Kruskal-Wallis test for significance. The results of the tests established at least one significant difference between the results of the performance windows in each of the outcome window categories, except for O24, using the *worst status* approach. For O24 no significant difference between the average class accuracies of the three performance window sizes was detected. Table 4 displays the results of the Dwass-Steel-Chritchlow-Fligner *post hoc* test for pair-wise comparisons.

Table 4: Average class accuracy *post hoc* analysis of Kruskal-Wallis test using Dwass-Steel-Chritchlow-Fligner. Results for the *worst status* (worst) label definition approach are provided. Note, no statistical significance was detected between the average class accuracies using O24. Statistical significance is indicated by *.

Insert Table 4 here

With the *worst status* approach, on each of the shorter outcome windows (O3, O6, and O12) statistically significant differences between P12 and both of the other performance windows (P6 and P18) were detected. This suggests that a 12-month performance window is the most suitable for similar sized, or shorter, outcome windows.

When the outcome window is extended to 18-months, statistically significant differences in the average class accuracies of P6 and P12 were detected. When taken with Figure 3, this suggests that when using an 18-month outcome window, a 6-month performance window is the least suitable. No significant difference existed between the

Insert Figure 4 here

Figure 4: Average class accuracy comparison of 3-month, 6-month, 12-month, 18-month, and 24-month outcome windows. *Worst status* label definition. The performance window is fixed at 12-months.

Insert Figure 5 here

Figure 5: Average class accuracy comparison of 3-month, 6-month, 12-month, 18-month, and 24-month outcome windows. *Current status* label definition approach. The performance window is fixed at 12-months.

average class accuracies of P12 and P18. This once again underlines the effects of seasonal trends and the lack of accumulated customer repayment behaviour transitions on classifier performance.

To summarise three separate performance window sizes were compared on each of the five outcome window sizes. Using the *worst status* label definition approach, we consider a 12-month performance window as best suited to the classification task - particularly when outcome window sizes of 3-months, 6-months, and 12-months were specified. For the longer outcome windows sizes, no statistical significance between the average class accuracies was detected to indicate the most suitable performance window. The average class accuracies of P12 were significantly better than its alternatives (P6 and P18) for three of the five outcome window sizes (O3, O6, and O12). As a result, P12 is carried forward into the next section as the fixed sized performance window with which to compare the different outcome window sizes.

4.2. Outcome Window Selection

In this section we quantify the effects outcome window size has on the average class accuracy of a classifier. In the literature there is no clear consensus on the most appropriate outcome window size. Figure 4 and Figure 5 compare classifier performance based on outcome window size and a fixed performance window size of 12-months using the *worst status* and *current status* approach, respectively.

In examining Figure 4, it is clear that a divide between the shorter outcome windows (O3, O6, and O12) and the longer outcome windows (O18, O24) exists. Also, the average class accuracies of the 3-month and 6-month outcome windows appear to be closely correlated.

Figure 5 reveals a clear ordering of classifier performance relative to the size of the outcome window. A logistic regression classifier using a 3-month outcome window consistently achieved the highest average class accuracy followed by, in order, logistic regression classifiers using O6, O18, O12, and O24. The relatively ineffectual performance of the logistic regression classifier using the longest outcome window, 24-months, seems to suggest that the transition from good to bad occurs over a short time period.

A Kruskal-Wallis test for significance on classifier performance in Figure 4 established at least one significant difference between the results. Similarly, a Kruskal-Wallis test for significance on classifier performance in Figure 5 established at least one significant difference between the average class accuracies. Table 5 displays the results of the Dwass-Steel-Chritchlow-Fligner *post hoc* test for pair-wise comparisons for both the *current status* and *worst status* approach.

Table 5: Average class accuracy, using a 12-month performance window, *post hoc* analysis of Kruskal-Wallis test using Dwass-Steel-Chritchlow-Fligner. Results for both label definition approaches: *worst status* (worst) and *current status* (current) are provided. Statistical significance is indicated by *.

Insert Table 5 here

Using the *worst status* approach, no statistical significance was detected between the performance of O3 and O6. However, statistically significant differences were found between the average class accuracies of both O3 and O6 when compared to those of O12, O18, and O24. Similarly, statistically significant differences were detected between the performance of the 12-month outcome window and that of both O18 and O24. These results indicate 3 groups - O3 and O6, which are better than O12, which, in turn, is better than O18 and O24. This suggests that, based on the relatively superior average class accuracies of O3 and O6, shorter outcome windows are best suited to a 12-month performance window for the current classification task. No statistical significance was detected between the average class accuracies of O18 and O24. It is worth noting that the performance of a classifier degrades substantially when the outcome window is extended beyond 12-months. For example, the difference in classifier performance between a 6-month and 12-month outcome window is relatively less than the difference in classifier performance between a 12-month and 18-month outcome window.

When analysing the *current status* approach, statistically significant differences between the average class accuracies of each outcome window were detected. Based on a 12-month performance window, this then enables us to infer the order of the most suitable outcome window size as follows: 3-months, 6-months, 18-months, 12-months, and 24-months. Note that the performance of a classifier degrades substantially once the outcome window is extended beyond 6-months. In contrast, for the *worst status* approach this degradation occurs when the outcome window is extended beyond 12-months.

To summarise, the results of the statistical comparison between the outcome windows clearly indicate that a logistic regression classifier using either a 3-month or 6-month outcome window, in conjunction with a 12-month performance window, results in a significantly higher average class accuracy than that of a classifier using a longer outcome window size. In addition, using the *current status* approach, the separation between the performance of the outcome windows is distinct compared to the *worst status* approach.

4.3. Current Status versus Worst Status

This section performs a direct comparison of the *current status* versus the *worst status* label definition approaches. The results of the comparison, performed using a 12-month performance window combined with each of the five outcome windows, are illustrated in Figure 6. It is clear from Figure 6 that a classification task based on the *worst status* approach and over longer outcome window sizes achieves a higher average class accuracy. However, the *current status* approach scored higher using the 3-month outcome window. A two-tailed Mann-Whitney U test was conducted for each outcome window to find statistically significant differences between the *current status* and *worst status* label definition approaches. For all outcome windows, except the 6-month outcome window, the two-tailed Mann-Whitney U test found statistically significant differences between the average class accuracies of the *current status* approach and *worst status* approach. This suggests that using *worst status* approach with an outcome window of 12-months or longer gives a higher assurance that the classification will be correct compared to the *current status* approach. Conversely, for a 3-month outcome window the *current status* approach presents an easier classification task than that of the *worst status* approach.

The *worst status* approach uses the entire outcome window and even though a customer may recover from arrears at some point during the outcome window, they are still classed as bad. This may be problematic as it implies that the relative rankings of default risk hold for the entirety of the outcome window (Thomas *et al.*, 2001). Conversely, with the *current status* approach, a customer who recovers from arrears during the outcome window is classed as good. As the *worst status* approach outperforms the *current status* approach over longer outcome windows this would suggest that once a customer's status changes they are unlikely to revert back to their original status. This has major implications for Irish mortgage holders, 9.2% of whom are currently in arrears of 3-months or more (Central Bank of Ireland, 2011).

Insert Figure 6a - 6e here

Figure 6: Average class accuracies (y-axis) of each outcome window. *Current status* versus *worst status*.

5. Conclusion

Behavioural scoring allows lenders to assess the likelihood of customers defaulting on their obligation during some specific outcome window. This assessment is based on customers' repayment behaviour over a fixed performance window. A customer's loan default status is defined either at the end of the outcome period (*current status*) or during the outcome period (*worst status*).

Using 7-years worth of data from the Irish market this article presented an extensive empirical evaluation of behavioural scoring models built using varying performance and outcome time horizons. In addition, the article also detailed an empirical comparison of both default status definition approaches.

Of the three separate performance windows used, the 12-month performance windows reported the highest average class accuracy on the shorter outcome window sizes. Over longer outcome window sizes the results exhibited greater ambiguity making it harder to identify an optimum performance window size. Frequently though, the 6-month performance window was unable to match the performance of the 12-month and 18-month performance windows. This relatively poor performance may be attributed to the effects of seasonality and the fact that less data is used to train the model.

The impact of outcome window size was examined using a 12-month performance window. Rather predictably, of the 5 five different outcome window sizes, the 3-month outcome window showed the best accuracy. Quite often the gap in performance between the 3-month outcome window and 6-month outcome window was statistically insignificant. As the length of the outcome windows increased, the differences in the average class accuracy of the various performance windows became less distinguishable.

The evaluation of the behavioural scoring models was conducted using two separate label definition approaches. The *worst status* approach performed better than the *current status* on a outcome window of 12-months or more. The *current status* approach was only superior when a 3-month outcome window was used. These results indicate that customers who fall into arrears are unlikely to recover.

Future work could concentrate on comparing the performance of classification models with duration models such as survival analysis. The work could also be expanded to identify customer accounts that settle early using survival analysis techniques. Furthermore, future work could assess the suitability of credit bureau data as a fundamental risk driver capable of determining defaults. Finally, an additional label definition approach which defines a bad based on a certain percentage of the arrears amount and the outstanding loan value could also be investigated.

References

Anderson, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, USA.

Andreeva, G. (2005). European generic scoring models using survival analysis. *Journal of the Operational research Society*, *57*, 1180–1187.

Andreeva, G., Ansell, J., and Crook, J. (2005). Modelling the purchase propensity: analysis of a revolving store card. *Journal of the Operational Research Society*, (pp. 1041–1050).

Baesens, B., Van Gestel, T., Stepanova, M., Van den Poel, D., and Vanthienen, J. (2005). Neural network survival analysis for personal loan data. *Journal of the Operational Research Society*, *56*, 1089–1098.

Banasik, J., Crook, J., and Thomas, L. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, (pp. 1185–1190).

Basel Committee on Banking Supervision (2006). *Intl Convergence of Capital Measurement and Capital Standards - A Revised Framework*. Basel II. Bank for Intl Settlements: Basel.

Bijak, K., and Thomas, L. (2011). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, .

Buchan, I. (2011). Statsdirect (version 2.7.8)[computer software]. Cheshire, United Kingdom: Statsdirect.

Burez, J., and den Poel, D. V. (2009). Handling class imbalance in customer churn prediction. *Expert Sys with Apps*, *36*, 4626–4636.

Central Bank of Ireland (2011). View Residential Mortgage Arrears and Repossession Statistics and Explanatory Notes. `http://www.centralbank.ie/press-area/press-releases/pages/residentialmortgagearrearsandrepossessionstatisticstodecember2011.aspx`.

Crone, S., and Finlay, S. (2011). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, .

DofE (2008). Latest House Prices, Loans and Profile of Borrowers Statistics. `http://www.environ.ie/en/Publications/StatisticsandRegularPublications/HousingStatistics/`. Accessed 3rd February 2011.

Finlay, S. (2010). Credit scoring for profitability objectives. *European Journal of Operational Research*, *202*, 528–537.

van Gestel, T., and Baesens, B. (2009). *Credit Risk Management: Basic Concepts*. Oxford University Press, USA.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, *11*, 10–18.

Hand, D. J., and Henley, W. E. (1997). Statistical classification methods in consumer credit scoring: a review. *J of the Royal Statistical Society. Series A (Statistics in Society)*, (pp. 523–541).

Hand, D. J., and Zhou, F. (2009). Evaluating models for classifying customers in retail banking collections. *J Opl Res Soc*, *Online*, doi:10.1057/jors.2009.129.

Lee, T., Chen, I. *et al.* (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, *28*, 743–752.

Malik, M., and Thomas, L. (2010). Modelling credit risk of portfolio of consumer loans. *J Opl Res Soc*, *61*, 411–420.

Mays, E. (2004). *Credit scoring for risk managers: the handbook for lenders*. Thomson South-Western Educational Publishing.

McDonald, R., Sturgess, M., Smith, K., Hawkins, M., and Huang, E. (2011). Non-linearity of scorecard log-odds. *International Journal of Forecasting*, .

McNab, H., and Wynn, A. (2000). *Principles and practice of consumer credit risk management*. Chartered Institute of Bankers and Institute of Financial Services and University of Manchester. Institute of Science and Technology.

Medema, L., Koning, R., and Lensink, R. (2009). A practical approach to validating a PD model. *J of Banking & Finance*, *33*, 701–708.

Sarlija, N., Bensic, M., and Zekic-Susac, M. (2008). Comparison procedure of predicting the time to default in behavioural scoring. *Expert Sys with Apps*, .

SAS (2009). A Comprehensive Credit Assessment Framework. `http://www.sas.com/resources/whitepaper/wp_4351.pdf`.

Sheskin, D. (1997). *Handbook of parametric and nonparametric statistical procedures*. CRC Press (Boca Raton, Fla.).

Stepanova, M., and Thomas, L. (2001). Phab scores: proportional hazards analysis behavioural scores. *Journal of the Operational Research Society*, (pp. 1007–1016).

Thomas, L. (2009). *Consumer credit models: pricing, profit, and portfolios*. Oxford University Press, USA.

Thomas, L., Ho, J., and Scherer, W. (2001). Time will tell: behavioural scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics*, *12*, 89–103.

Thomas, L. C., Edelman, D. B., and Crook, J. N. (2002). *Credit scoring and its applications*. Society for Industrial Math.

Till, R., and Hand, D. (2003). Behavioural models of credit card usage. *Journal of Applied Statistics*, *30*, 1201–1220.

Tsai, C. F. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, *22*, 120–127.

Wolfe, D., and Hollander, M. (1999). *Nonparametric statistical methods, 2nd Edition.*. John Wiley New York.

Figure 1 caption: Behavioural scoring performance window and outcome window.
Figure 2 caption: Experiment set-up for a 12 month performance window and a 3-month outcome window (Out).
Figure 3 caption: Average class accuracies (*y*-axis) of each outcome window. *Worst status* label definition.
  Sub-Figure 3a caption: 3-Month outcome window
  Sub-Figure 3b caption: 6-Month outcome window
  Sub-Figure 3c caption: 12-Month outcome window
  Sub-Figure 3d caption: 18-Month outcome window
  Sub-Figure 3e caption: 24-Month outcome window
Figure 4 caption: Average class accuracy comparison of 3-month, 6-month, 12-month, 18-month, and 24-month outcome windows. *Worst status* label definition. The performance window is fixed at 12-months.
Figure 5 caption: Average class accuracy comparison of 3-month, 6-month, 12-month, 18-month, and 24-month outcome windows. *Current status* label definition approach. The performance window is fixed at 12-months.
Figure 6 caption: Average class accuracies (*y*-axis) of each outcome window. *Current status* versus *worst status*.
  Sub-Figure 6a caption: 3-Month outcome window
  Sub-Figure 6b caption: 6-Month outcome window
  Sub-Figure 6c caption: 12-Month outcome window
  Sub-Figure 6d caption: 18-Month outcome window
  Sub-Figure 6e caption: 24-Month outcome window

Table 1: Behavioural scoring data sources and associated features (McNab and Wynn, 2000).

Table 2: ICB data features. Features removed from the ICB data during data preparation are indicated by *.

Table 3: Combination features generated based on the arrears repayment indicator. This example is based on a 18-month performance window.

Table 4: Average class accuracy *post hoc* analysis of Kruskal-Wallis test using Dwass-Steel-Chritchlow-Fligner. Results for the *worst status* (worst) label definition approach are provided. Note, no statistical significance was detected between the average class accuracies using O24. Statistical significance is indicated by *.

Table 5: Average class accuracy, using a 12-month performance window, *post hoc* analysis of Kruskal-Wallis test using Dwass-Steel-Chritchlow-Fligner. Results for both label definition approaches: *worst status* (worst) and *current status* (current) are provided. Statistical significance is indicated by *.

***Highlights (for review)**

Highlights

- We perform an empirical evaluation of behavioural scoring models.
- We examine performance and outcome window sizes, default status definitions.
- A 12-month performance window reported the highest average class accuracy.
- Relatively minor differences in the results of 3-month and 6-month outcome windows.
- Current status is more accurate than worst status label definition approach.

Table 1

Data Source	Feature Example
Delinquency history	Ever in arrears Maximum arrears level
Usage history	Balance-to-limit ratio Balance trends
Static information	Customer age Application score
Payment/purchase history	Purchase frequency Type of retail goods purchased
Collections activity	Outcomes Contact frequency
Revolving credit transactions	Number Type (retail/cash)
Customer service contacts	Inbound contact Outbound contact
Promotions history	Number of offers Outcome of offers
Bureau data	Generic scores Shared account information

Table 2

Type	Feature Name	Description
Application Data	Account opening date*	Date of loan drawdown
	Loan amount	Total repayment amount
	Term of loan	Length of loan in years
	Loan protection	Financial insurance
	Customer location	Current residence of the customer
	Account association	Single or joint account
	Payment frequency	Monthly or fortnightly repayment
	Date of birth*	Customer date of birth
Behavioural data	Account update date*	Date of loan update
	Loan installment amount	Amount repaid for update
	Outstanding loan balance	Overall amount owed
	Vintage	Age of loan in months
	Repayment indicators	Monthly loan status

Table 3

Feature Name	Description
Arrears ever	Have arrears occurred at any stage during the performance window
Arrears 1-to-3	Have arrears occurred in the last 3 months
Arrears 4-to-6	Have arrears occurred in months 4-to-6
Arrears 7-to-12	Have arrears occurred in months 7-to-12
Arrears 13-to-18	Have arrears occurred in months 13-to-18
Arrears worst	Highest number of missed repayments
Missed single	Only one missed repayment on a single occasion
Missed twice	Only one missed repayment on two separate occasions
Missed thrice	Only one missed repayment on three separate occasions
Missed multi	Only one missed repayment on more than three separate occasions
Missed double	At most 2 repayments in arrears
Missed treble	At most 3 repayments in arrears
Missed cont > 3	At most 3 or more repayments in arrears
Arrears closed	Was the account in arrears before it was closed
One current	Has the account gone directly from 1 missed repayment to normal
Two current	Has the account gone directly from 2 missed repayments to normal
Three current	Has the account gone directly from 3 missed repayments to normal

Table 4

Outcome	Performance	*p*-value
O3	P18 vs. P12	0.0006*
	P18 vs. P6	0.0001*
	P12 vs. P6	< 0.0001*
O6	P18 vs. P12	0.0003*
	P18 vs. P6	0.5318
	P12 vs. P6	< 0.0001*
O12	P18 vs. P12	0.0041*
	P18 vs. P6	0.9919
	P12 vs. P6	0.0003*
O18	P18 vs. P12	0.4409
	P18 vs. P6	0.6139
	P12 vs. P6	0.0085*

Table 5

Outcome	*p*-value	
	worst	**current**
O3 vs. O6	0.9766	0.0001*
O3 vs. O12	0.0052*	< 0.0001*
O3 vs. O18	< 0.0001*	< 0.0001*
O3 vs. O24	< 0.0001*	< 0.0001*
O6 vs. O12	0.0369*	< 0.0001*
O6 vs. O18	< 0.0001*	0.0002*
O6 vs. O24	< 0.0001*	< 0.0001*
O12 vs. O18	0.0006*	0.0394*
O12 vs. O24	< 0.0001*	0.0263*
O18 vs. O24	0.8156	0.0011*

Figure 1

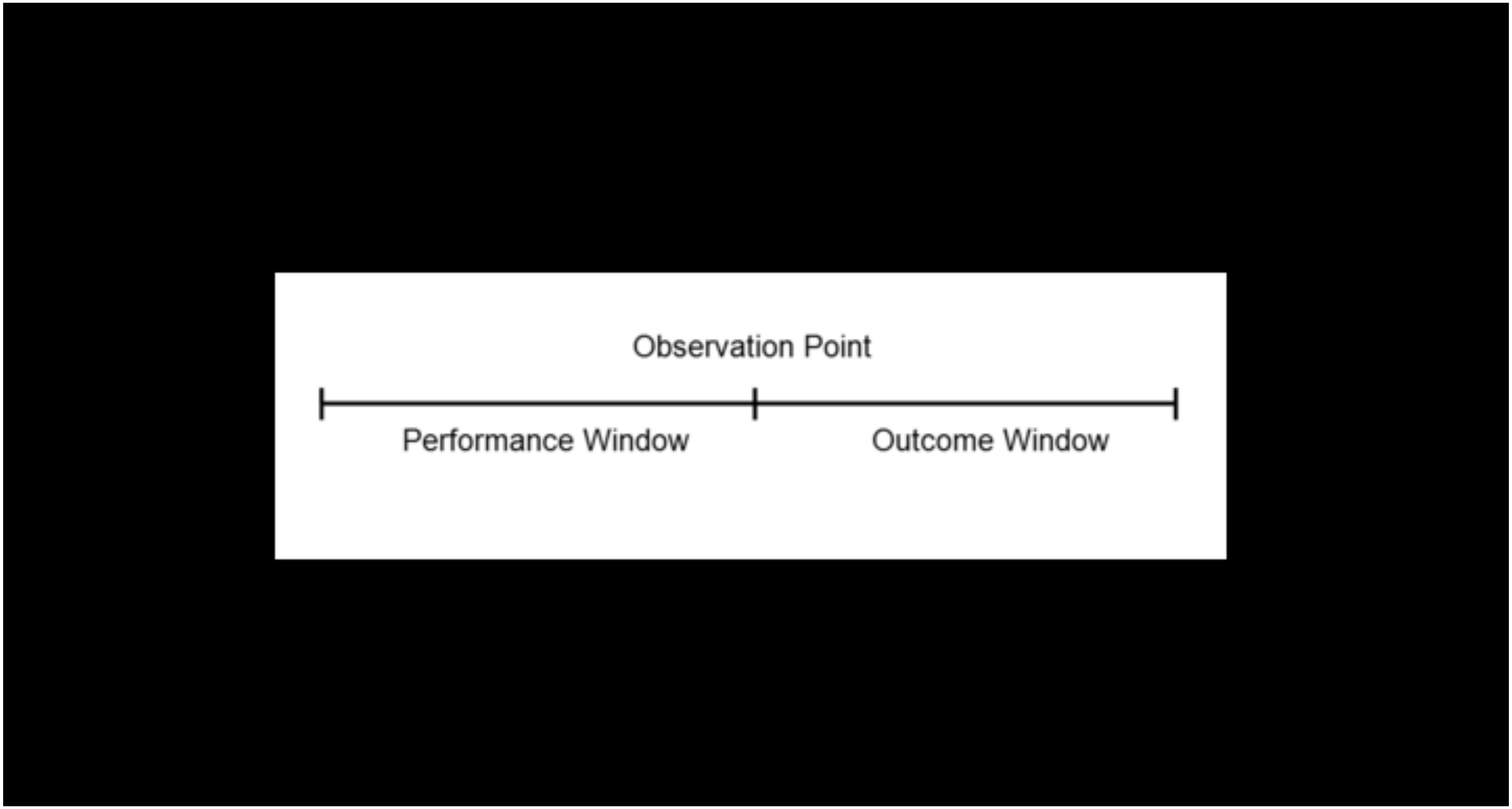

Observation Point
Performance Window
Outcome Window

Figure 2

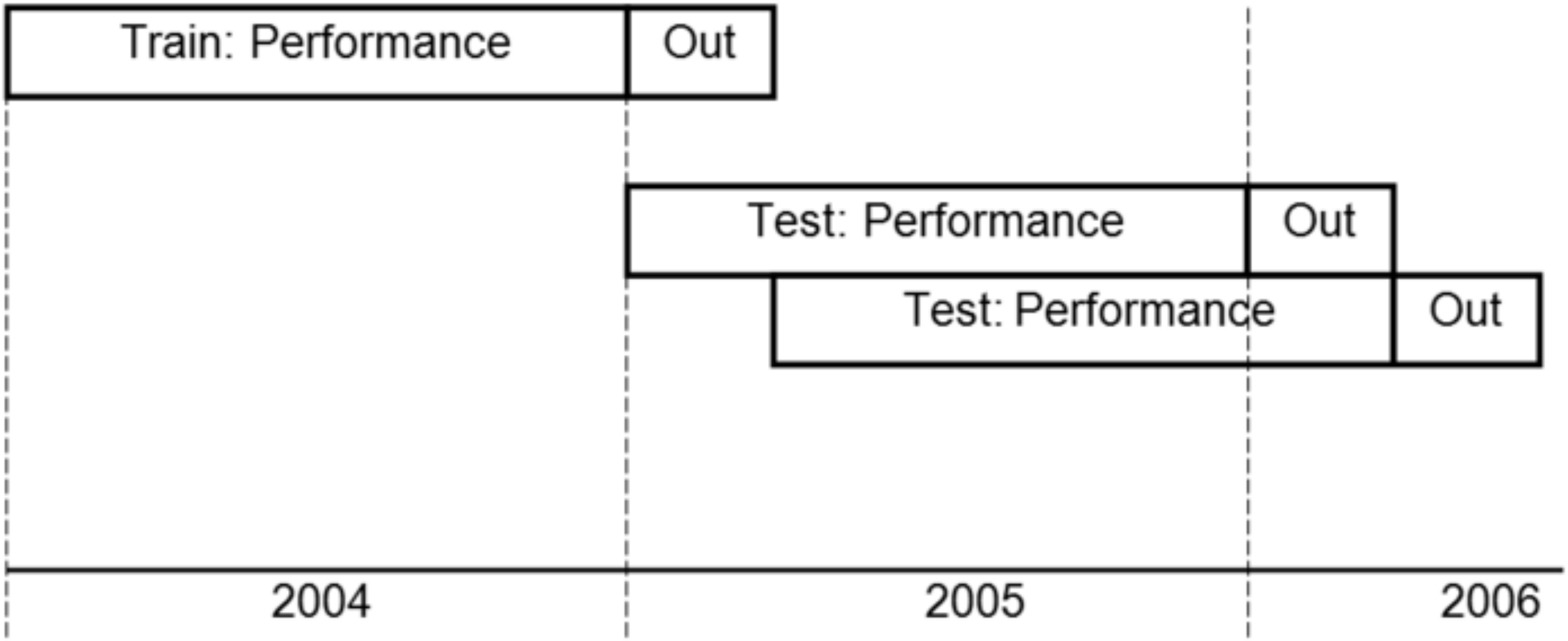

Train: Performance
Out
Test: Performance
Out
Test: Performance
Out
2004
2005
2006

Figure 3a

Click here to download high resolution image

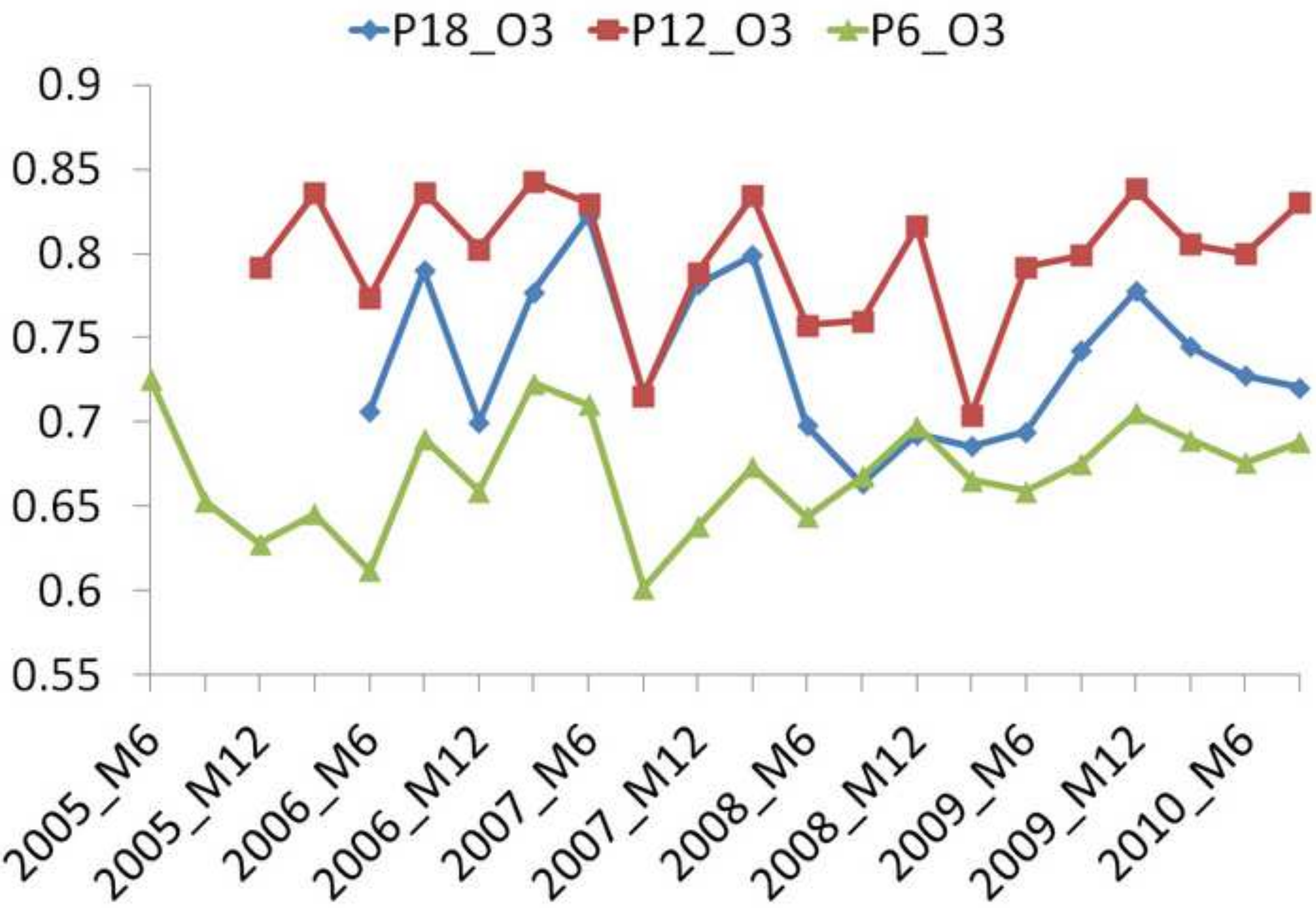


Figure 3b
Click here to download high resolution image

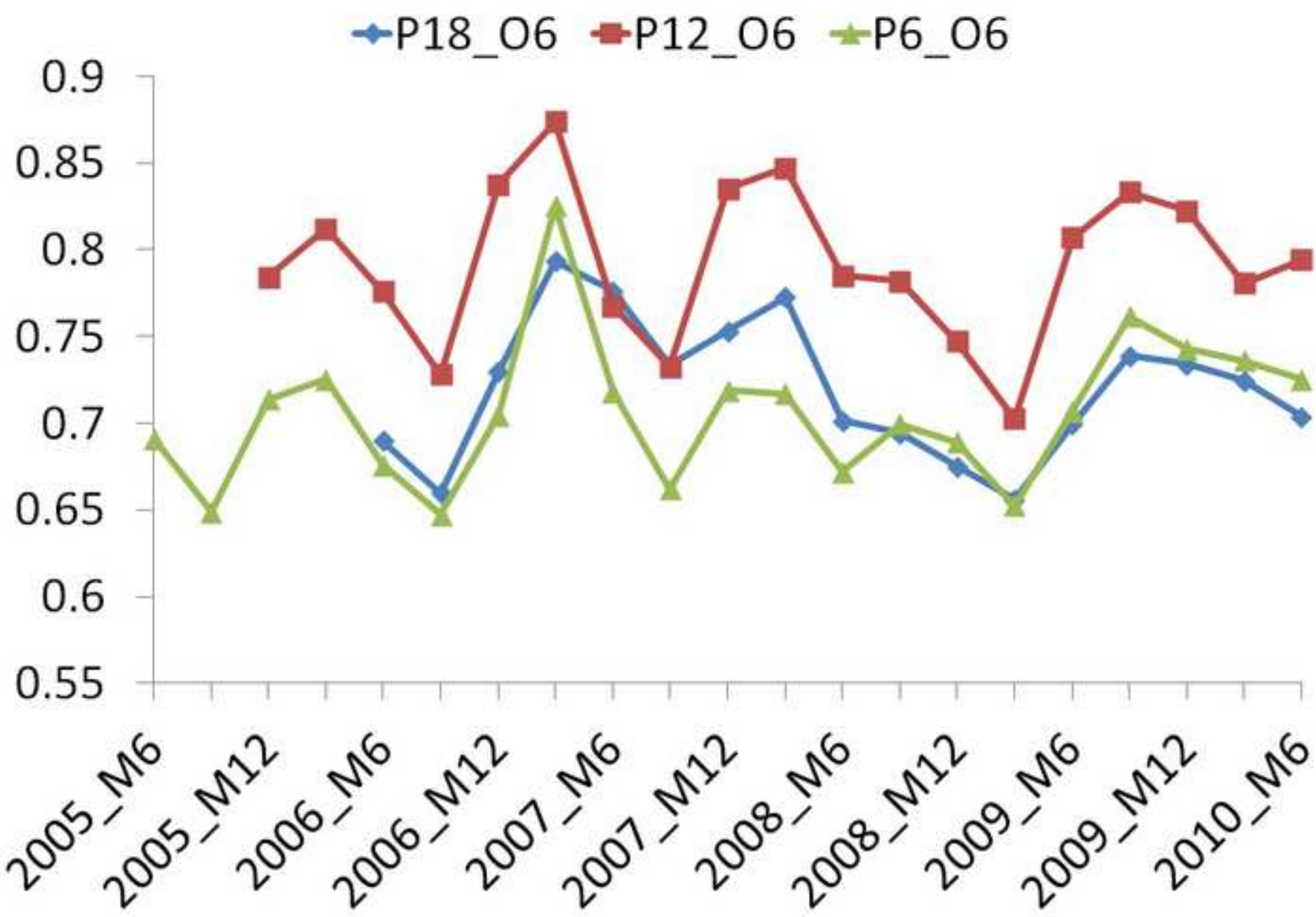


Figure 3c

Click here to download high resolution image

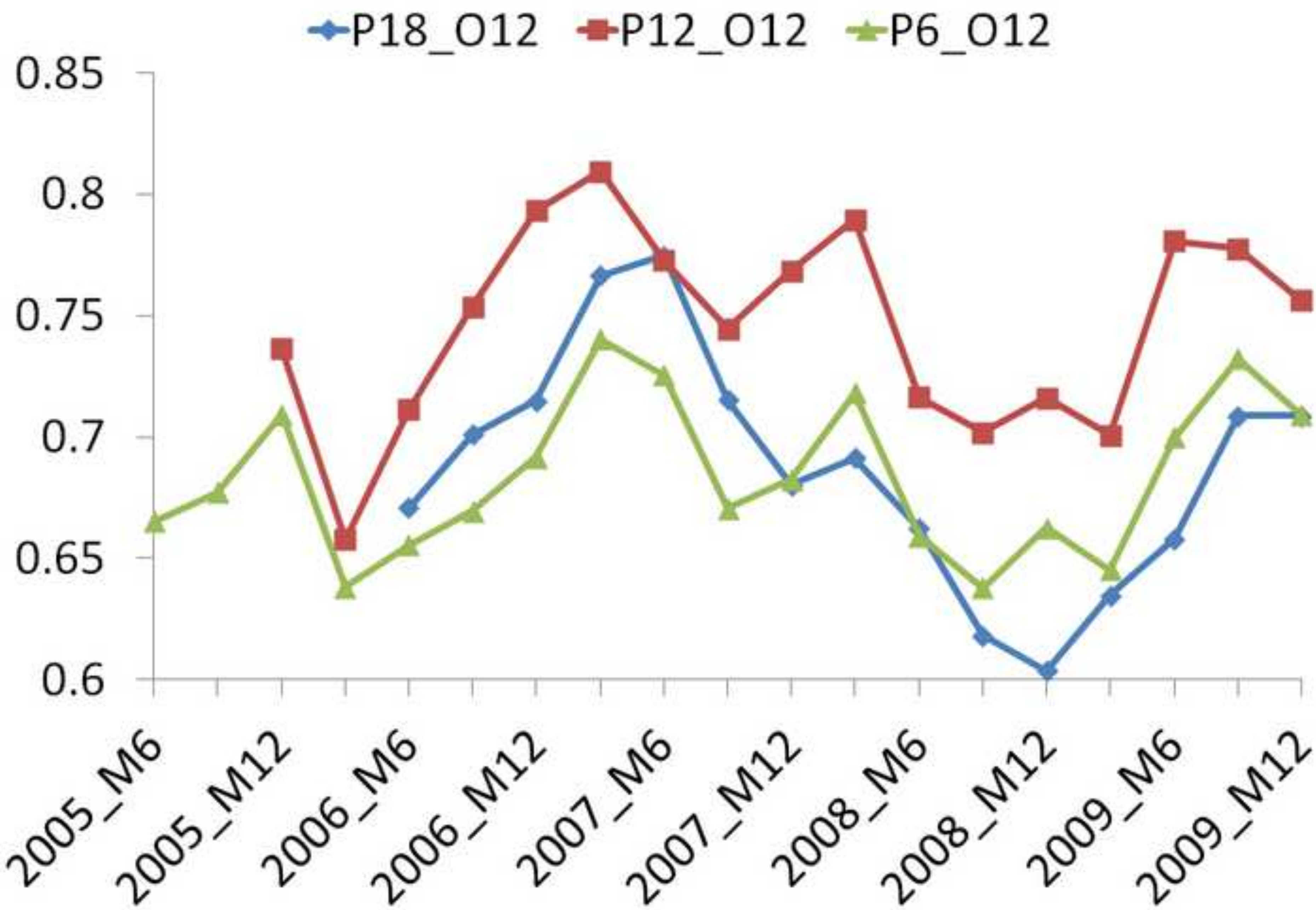


Figure 3d

Click here to download high resolution image

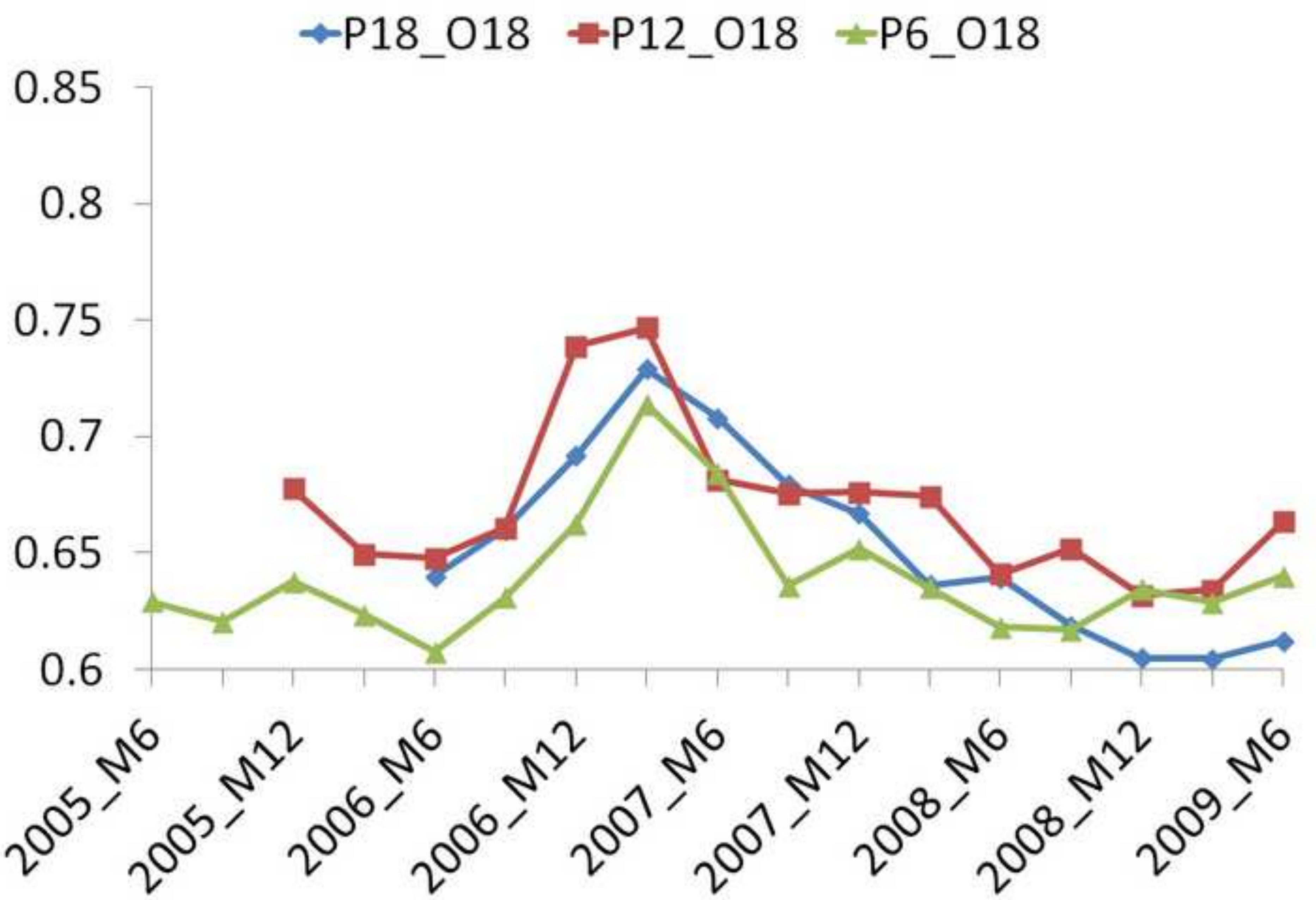


Figure 3e

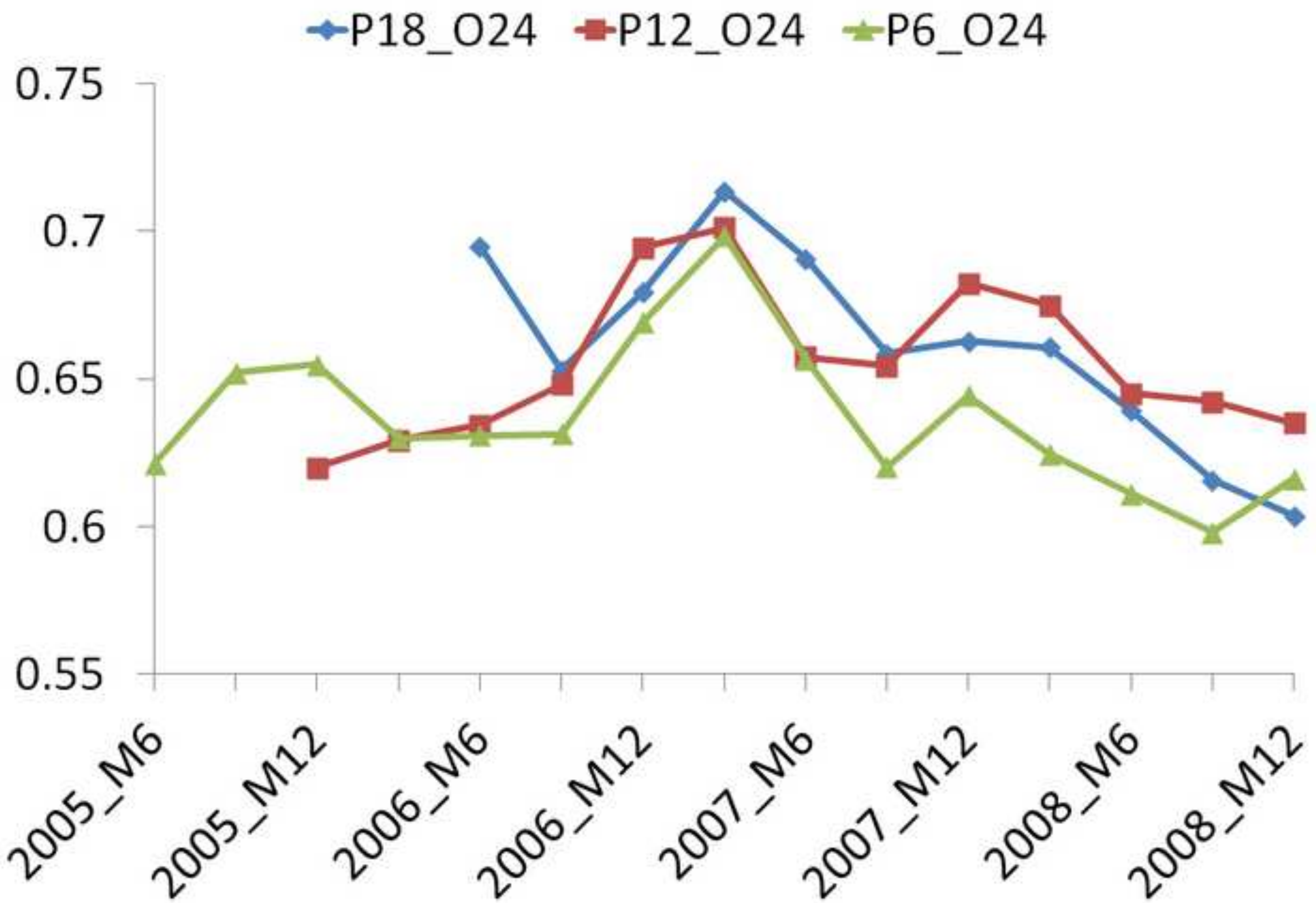

P18_O24
P12_O24
P6_O24
0.75
0.7
0.65
0.6
0.55
2005_M6
2005_M12
2006_M6
2006_M12
2007_M6
2007_M12
2008_M6
2008_M12

Figure 4

Click here to download high resolution image

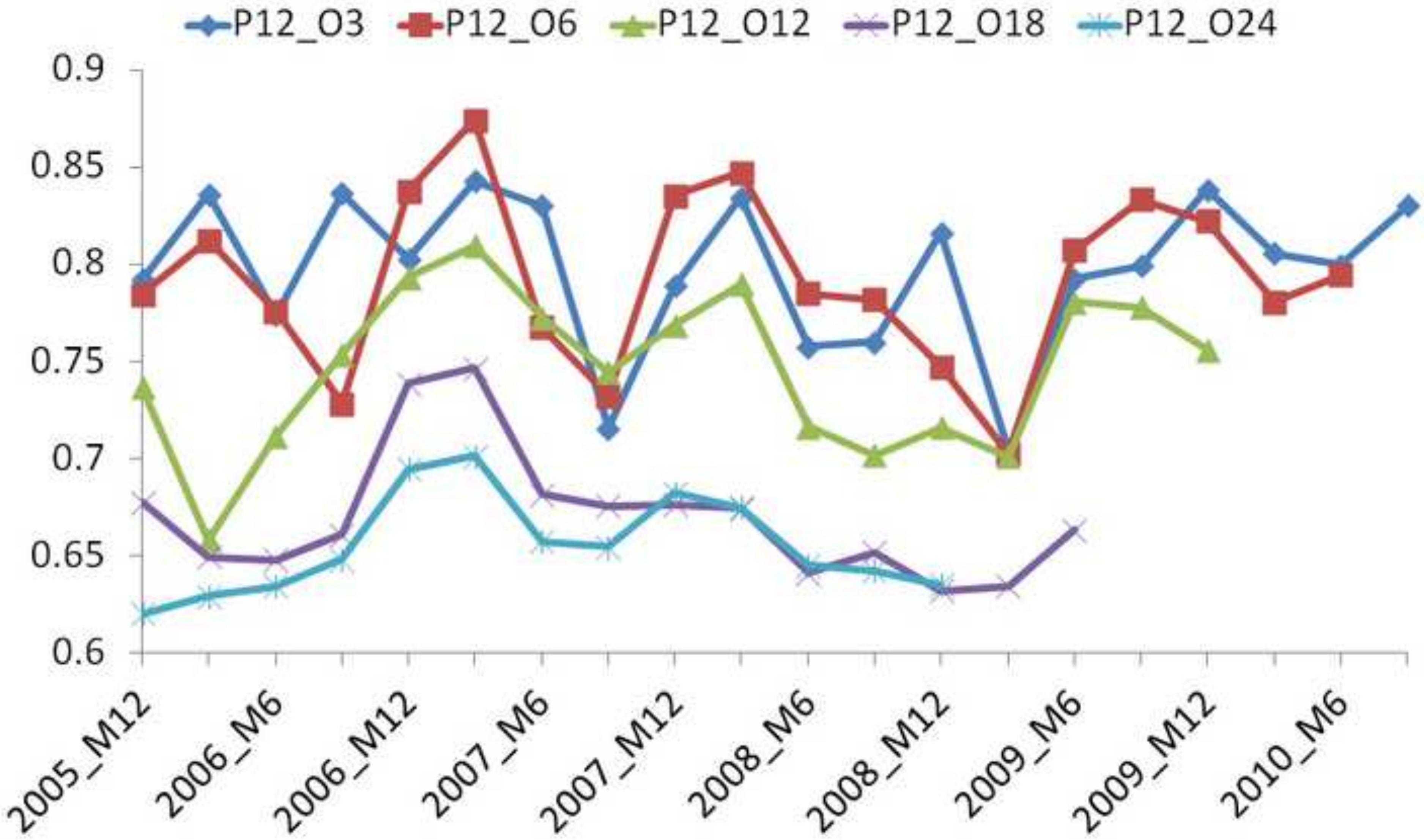


Figure 5

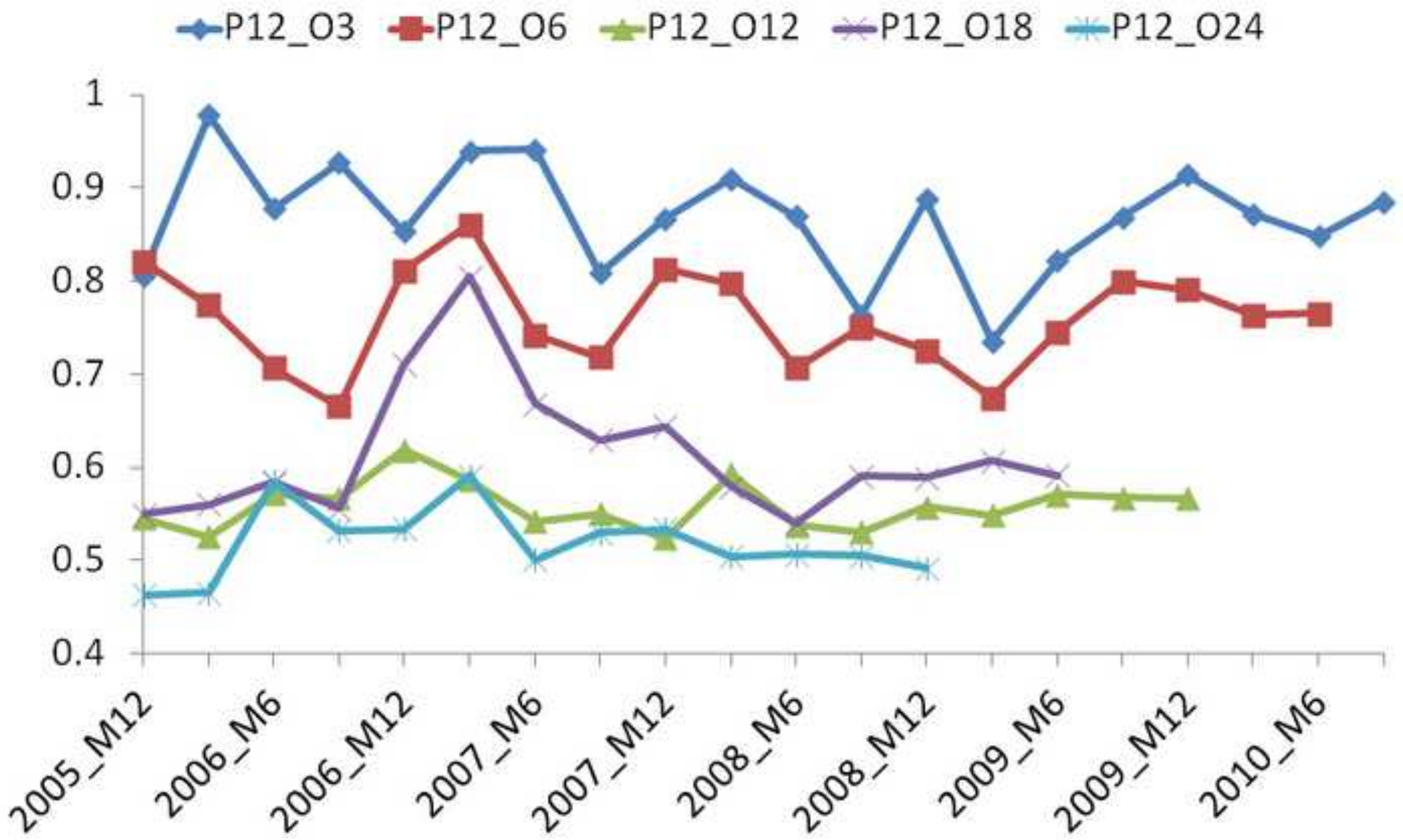

P12_O3
P12_O6
P12_O12
P12_O18
P12_O24
1
0.9
0.8
0.7
0.6
0.5
0.4
2005_M12
2006_M6
2006_M12
2007_M6
2007_M12
2008_M6
2008_M12
2009_M6
2009_M12
2010_M6

Figure 6a

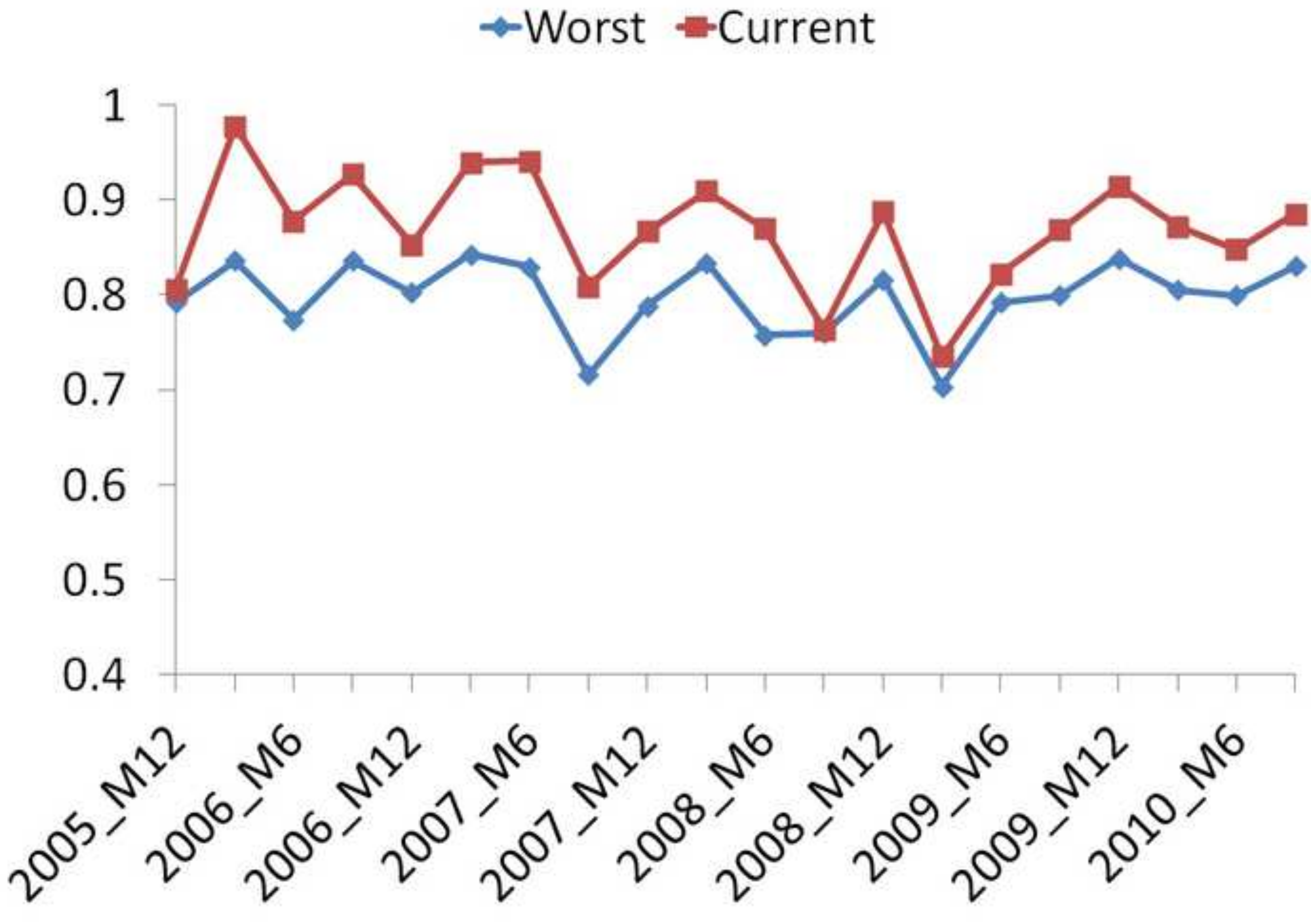

Worst
Current
1
0.9
0.8
0.7
0.6
0.5
0.4
2005_M12
2006_M6
2006_M12
2007_M6
2007_M12
2008_M6
2008_M12
2009_M6
2009_M12
2010_M6

Figure 6b

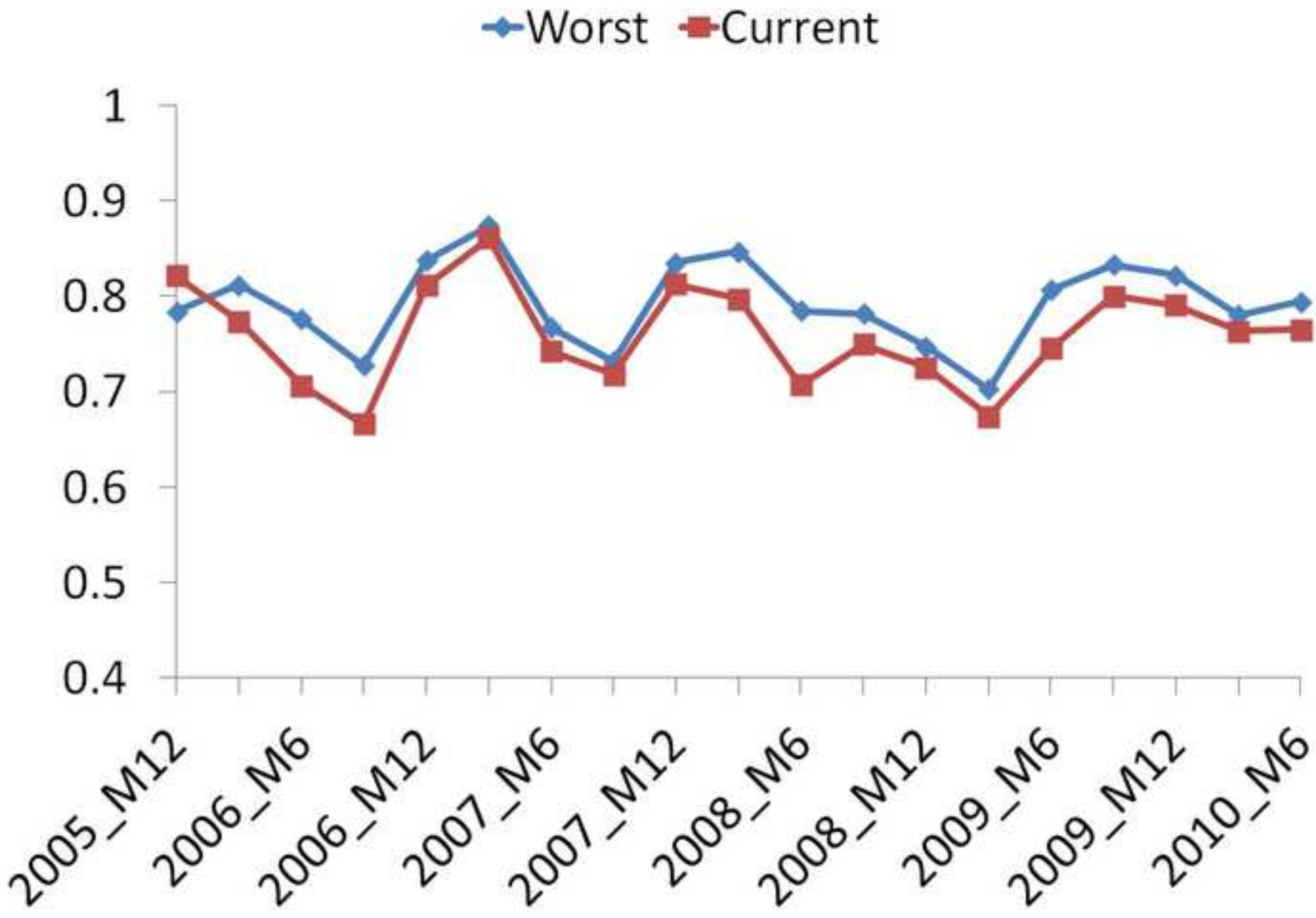
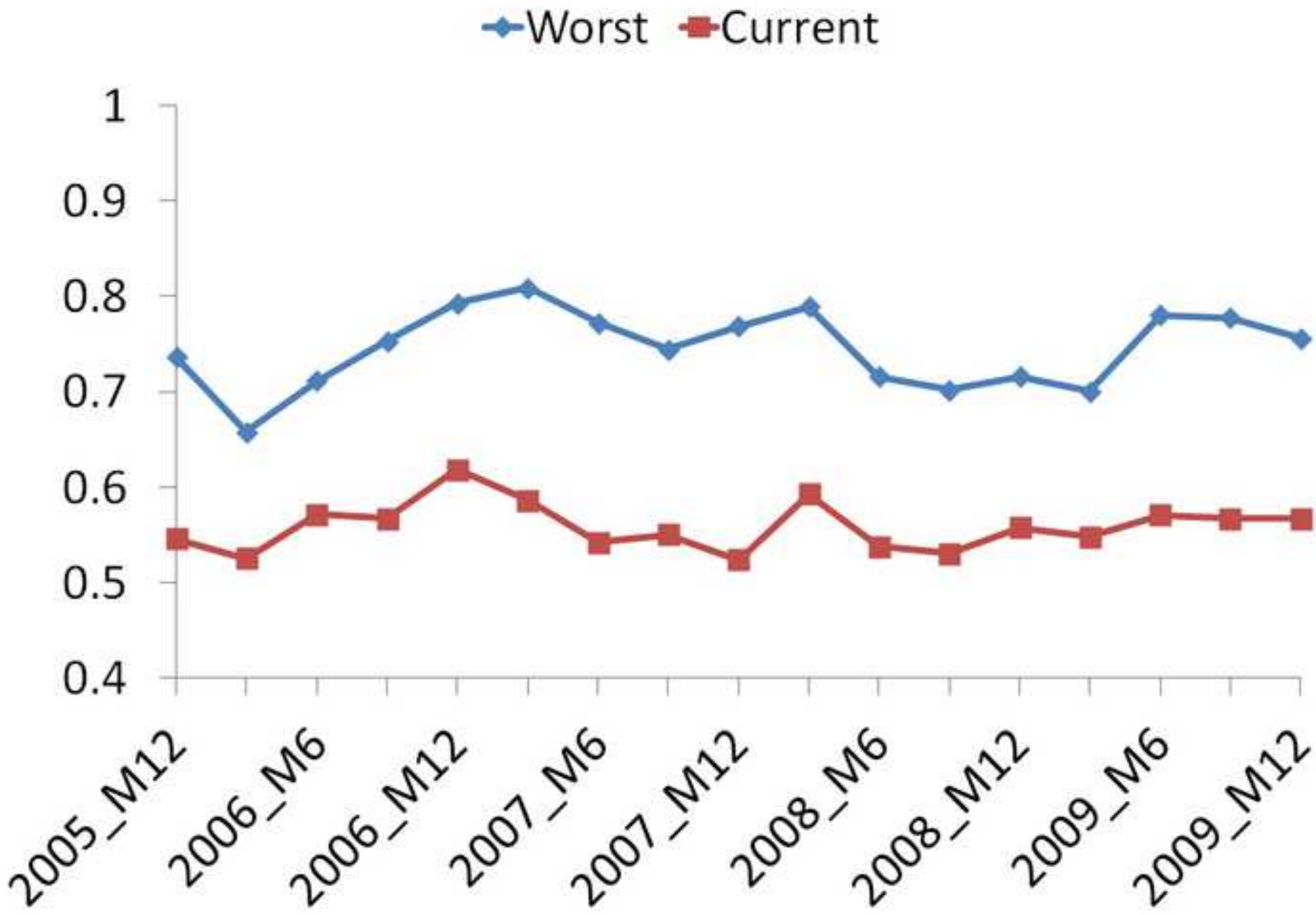

Worst
Current
1
0.9
0.8
0.7
0.6
0.5
0.4
2005_M12
2006_M6
2006_M12
2007_M6
2007_M12
2008_M6
2008_M12
2009_M6
2009_M12
2010_M6

Figure 6c



Worst
Current
1
0.9
0.8
0.7
0.6
0.5
0.4
2005_M12
2006_M6
2006_M12
2007_M6
2007_M12
2008_M6
2008_M12
2009_M6
2009_M12

Figure 6d

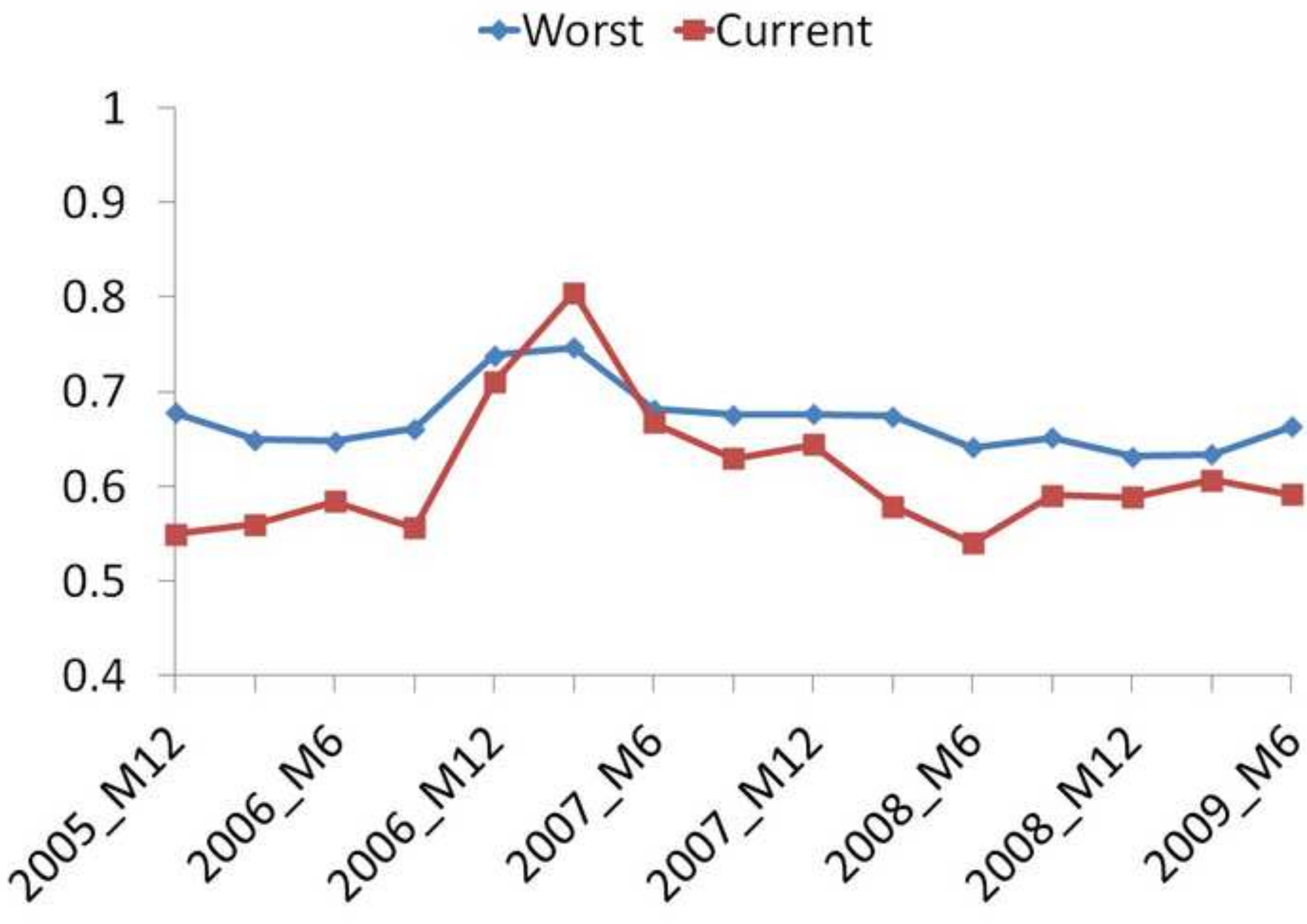

Worst
Current
1
0.9
0.8
0.7
0.6
0.5
0.4
2005_M12
2006_M6
2006_M12
2007_M6
2007_M12
2008_M6
2008_M12
2009_M6

Figure 6e

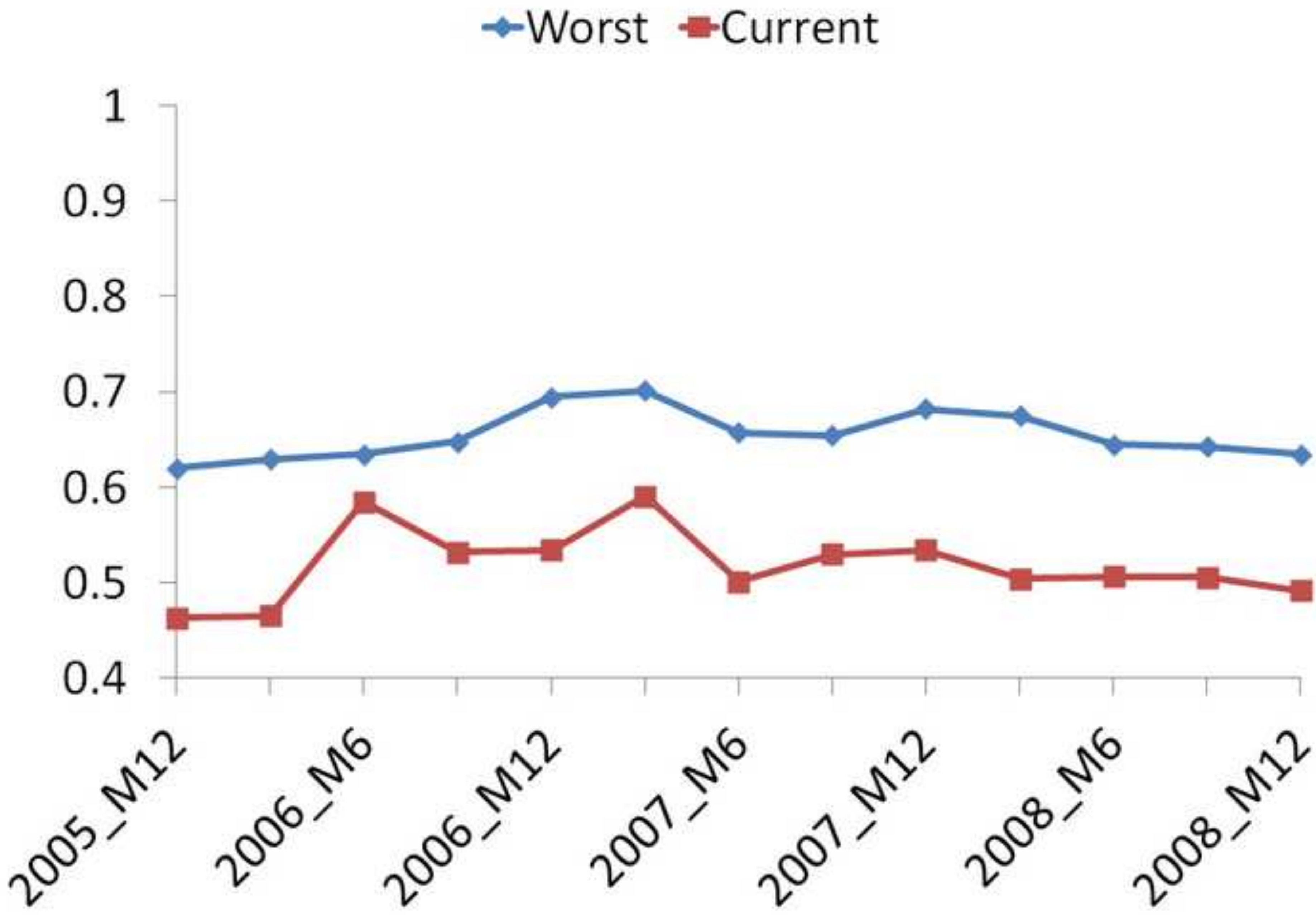

Worst
Current
1
0.9
0.8
0.7
0.6
0.5
0.4
2005_M12
2006_M6
2006_M12
2007_M6
2007_M12
2008_M6
2008_M12

Figure 3a greyscale

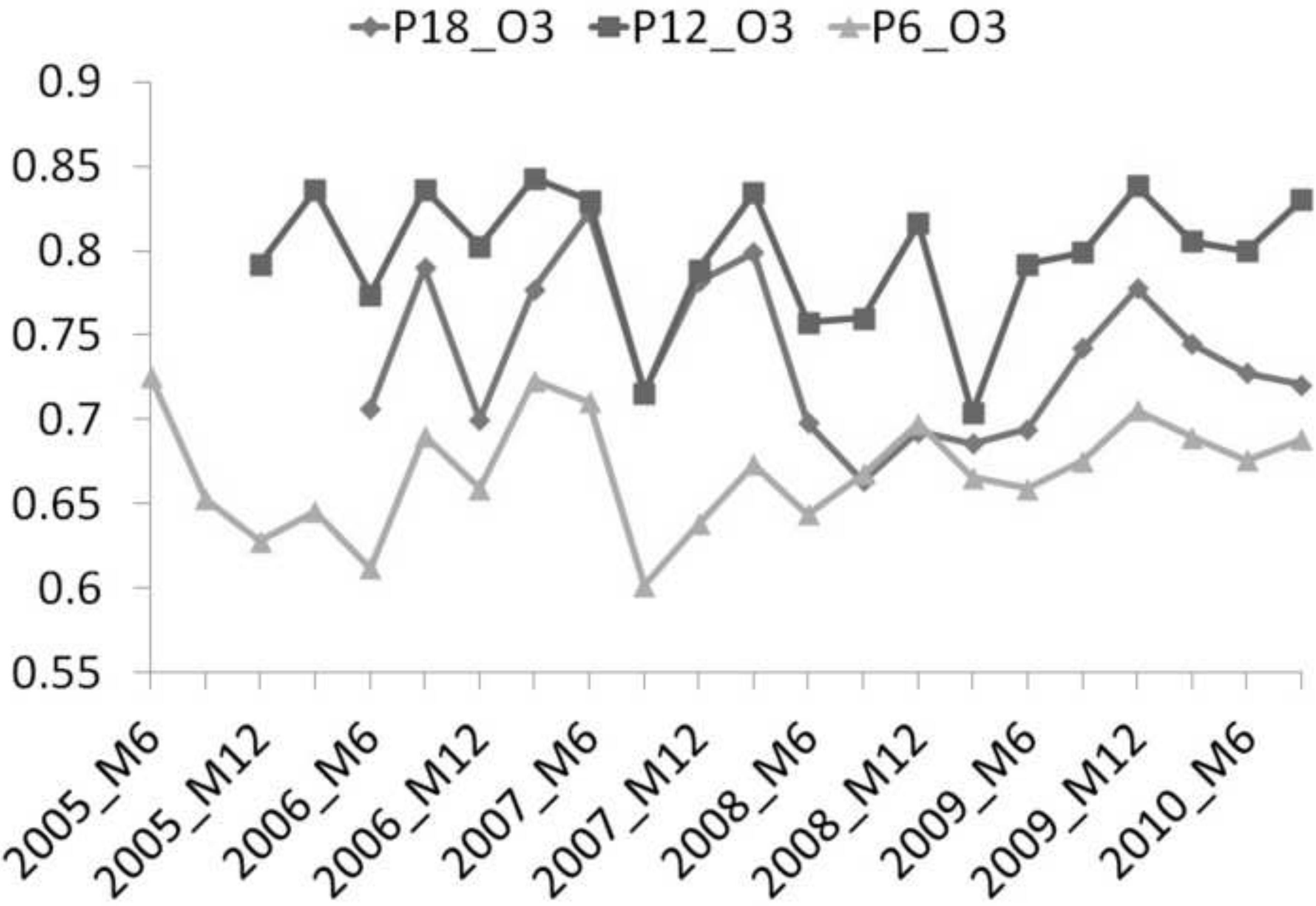


Figure 3b greyscale

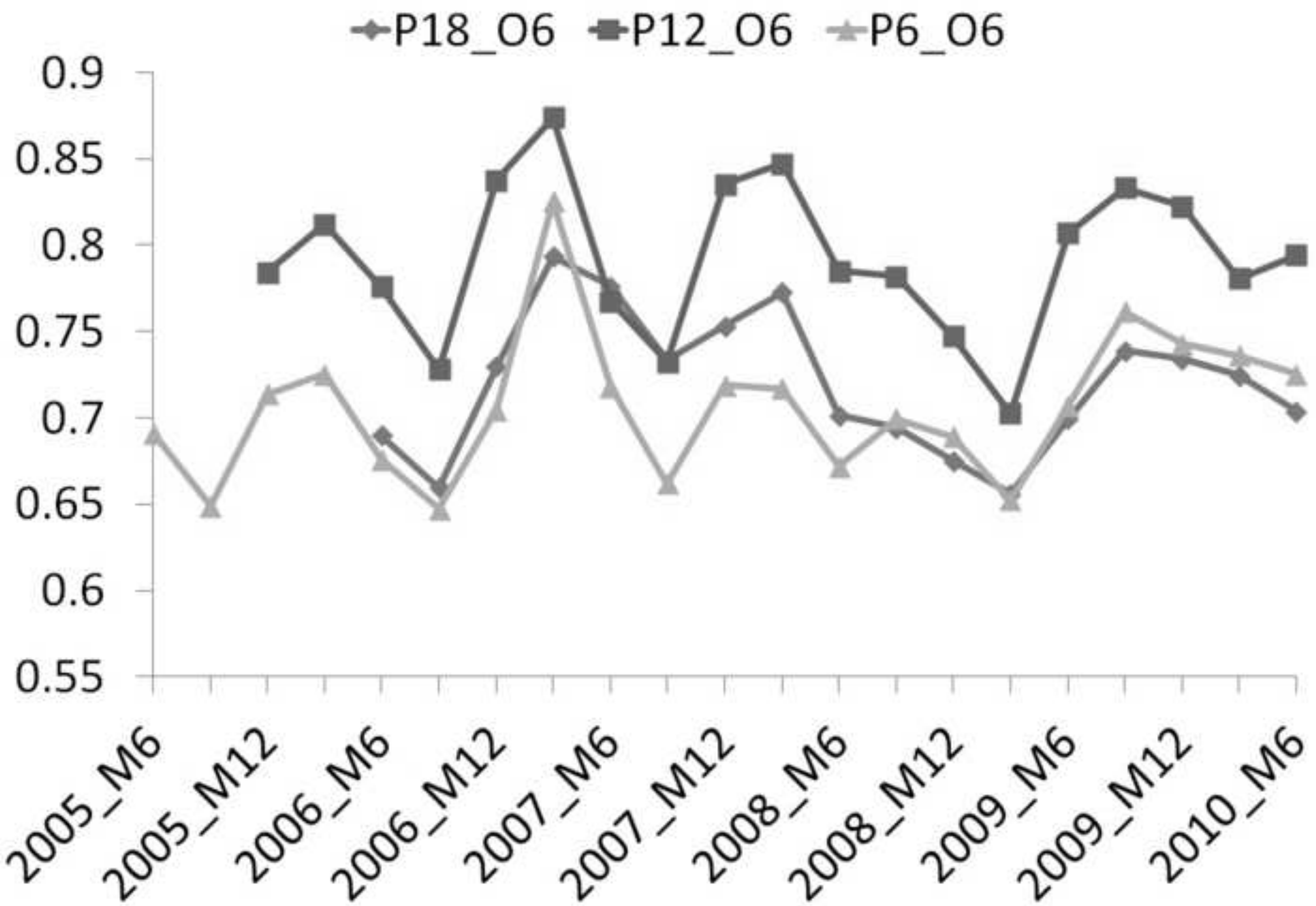


Figure 3c greyscale

Click here to download high resolution image

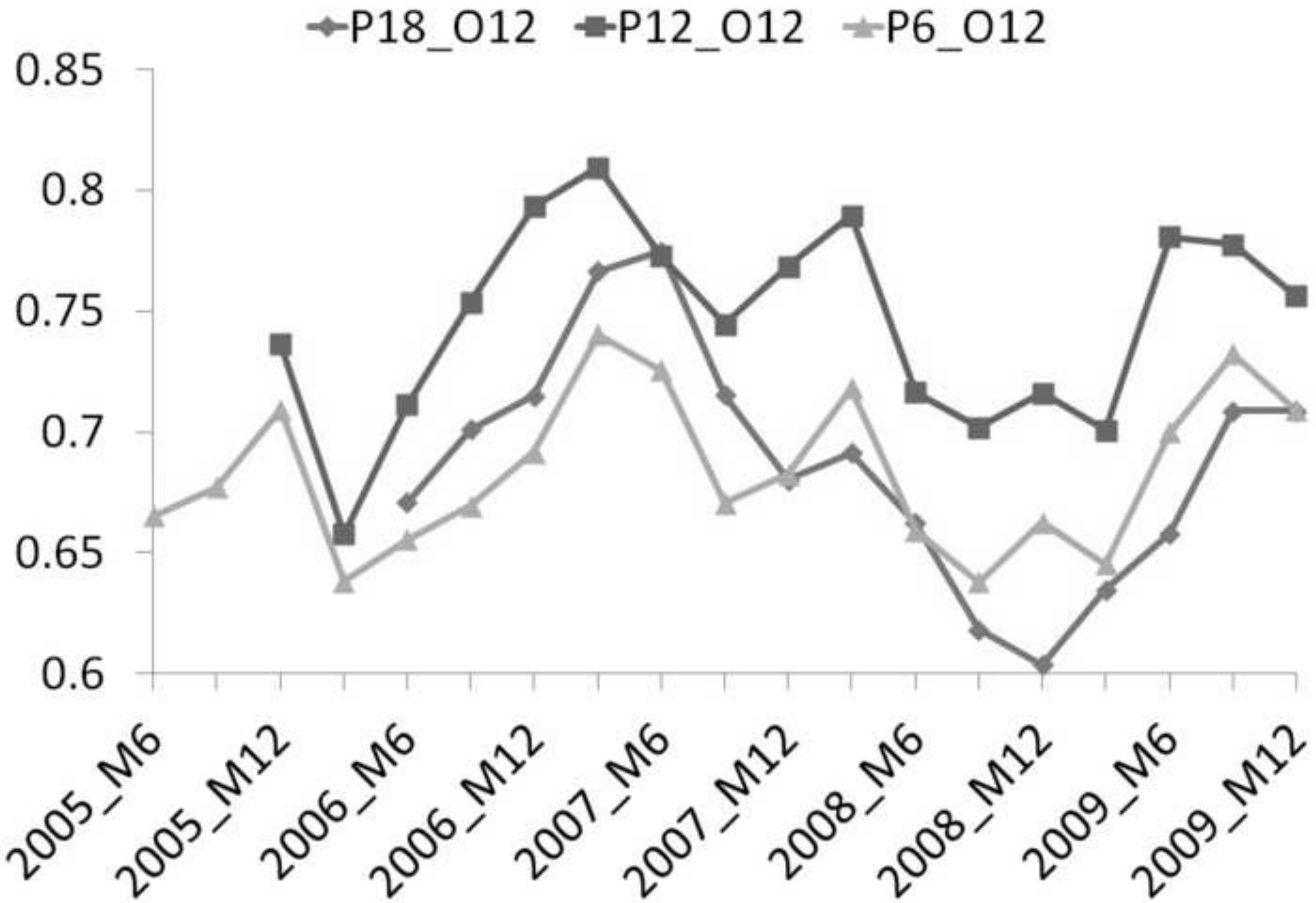


Figure 3d greyscale

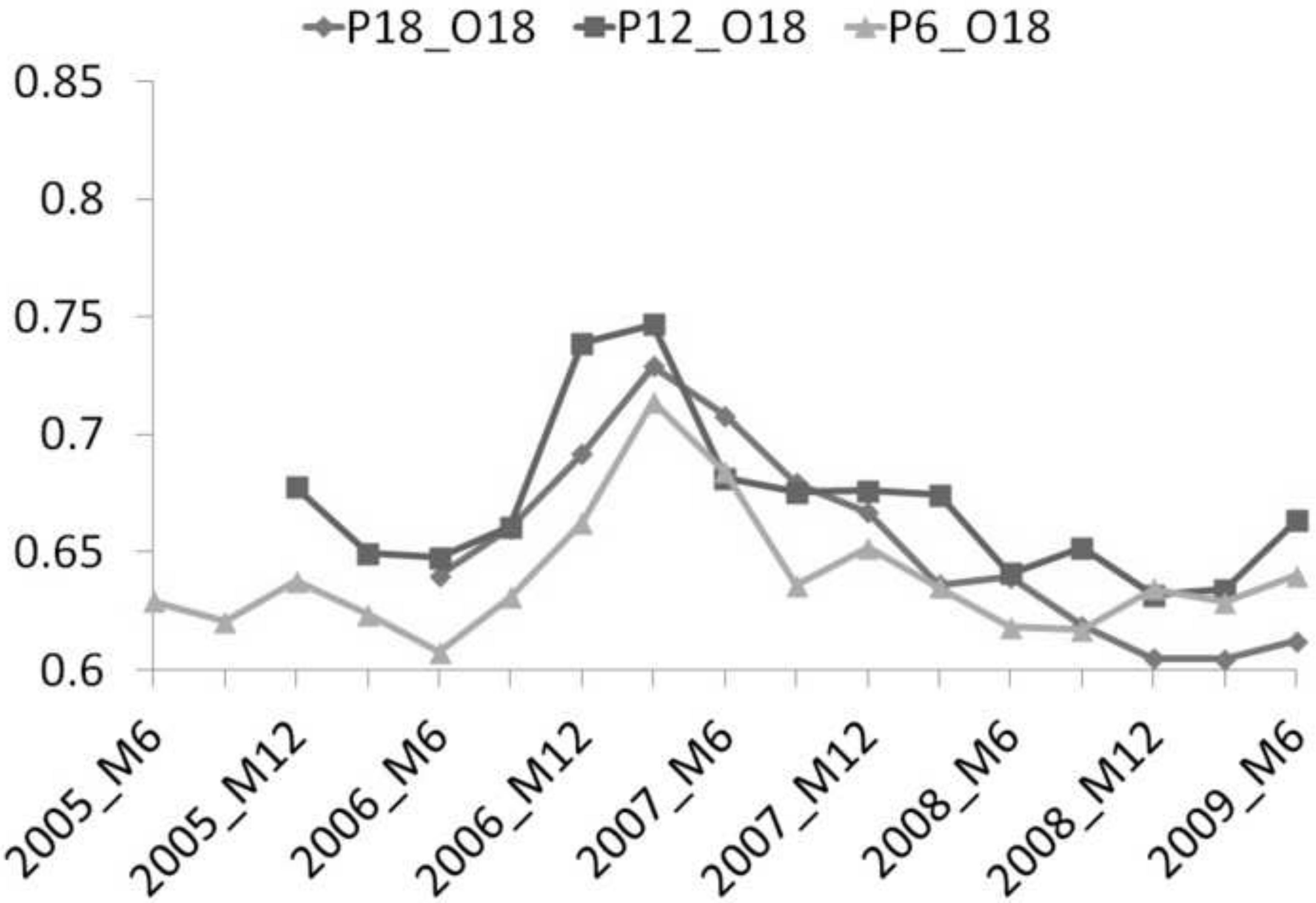


Figure 3e greyscale

Click here to download high resolution image

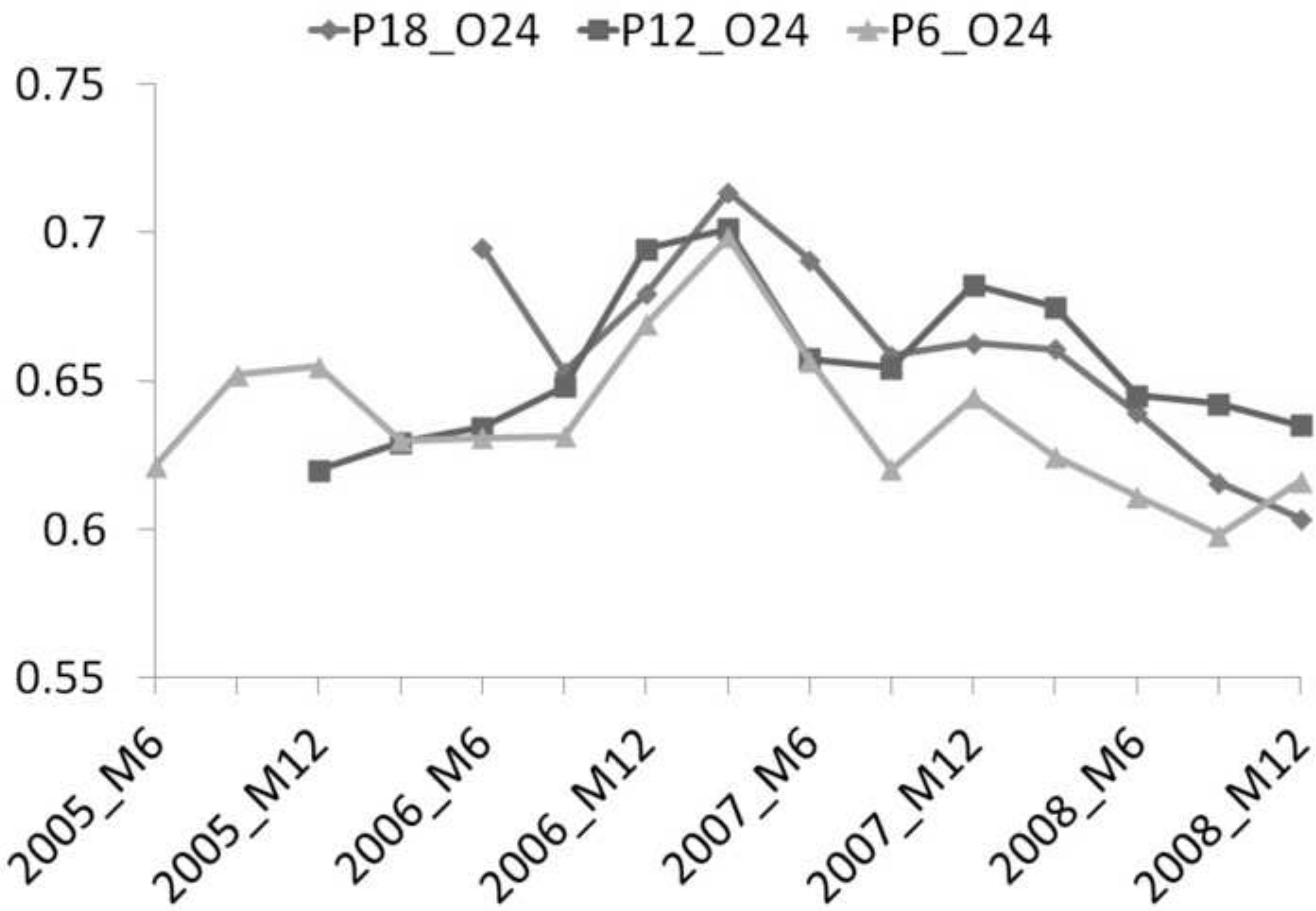


Figure 4 greyscale

Click here to download high resolution image

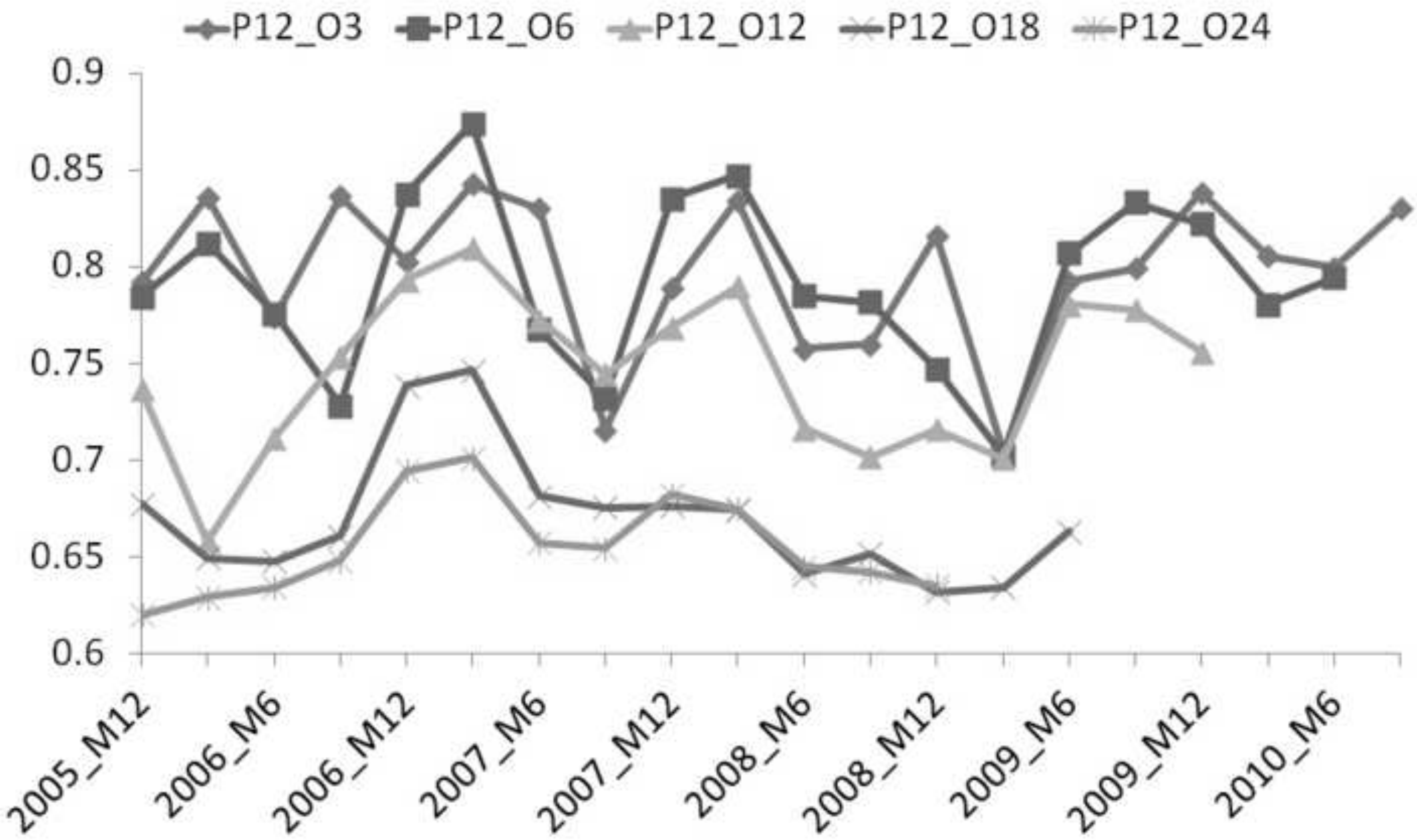


Figure 5 greyscale

Click here to download high resolution image

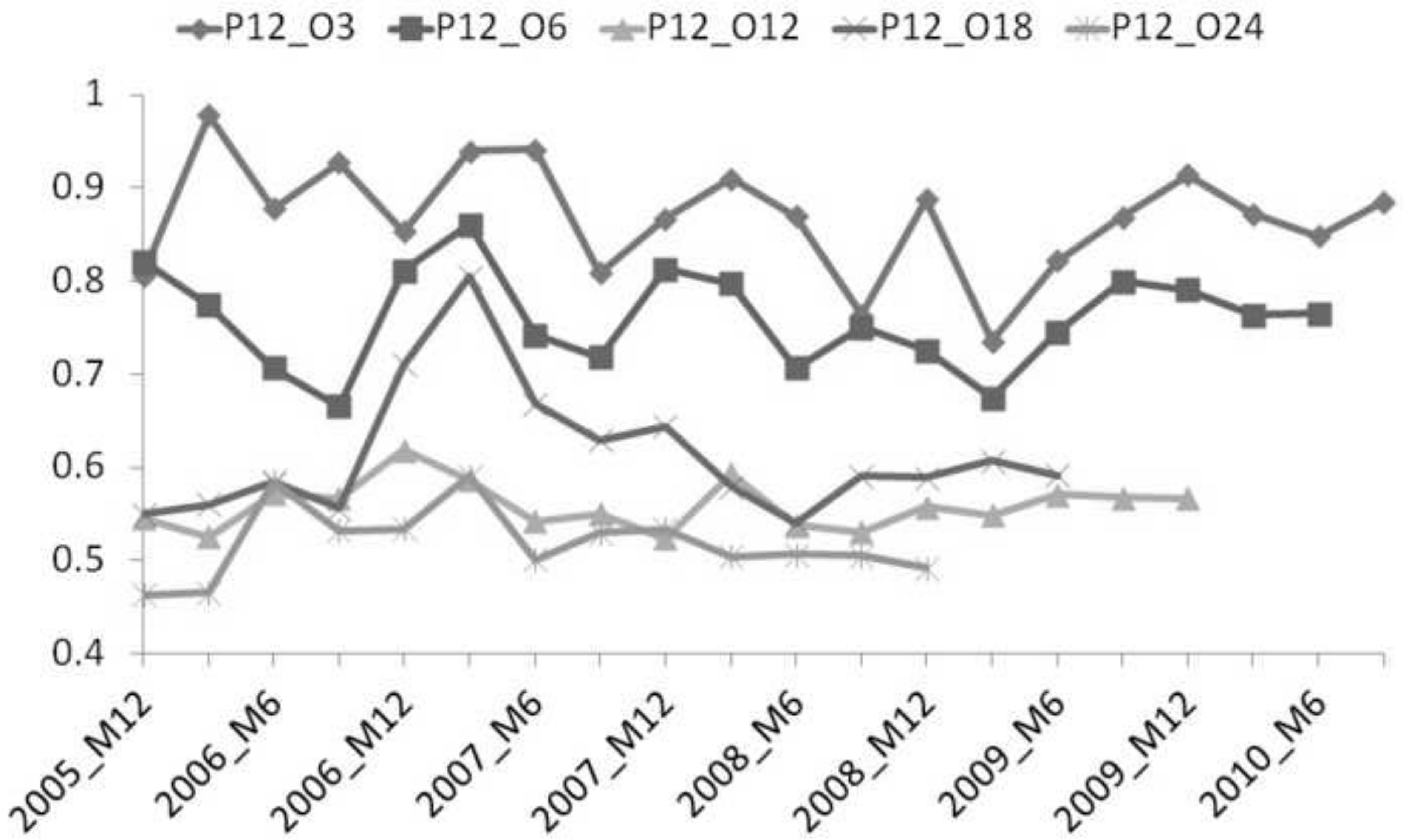


Figure 6a greyscale

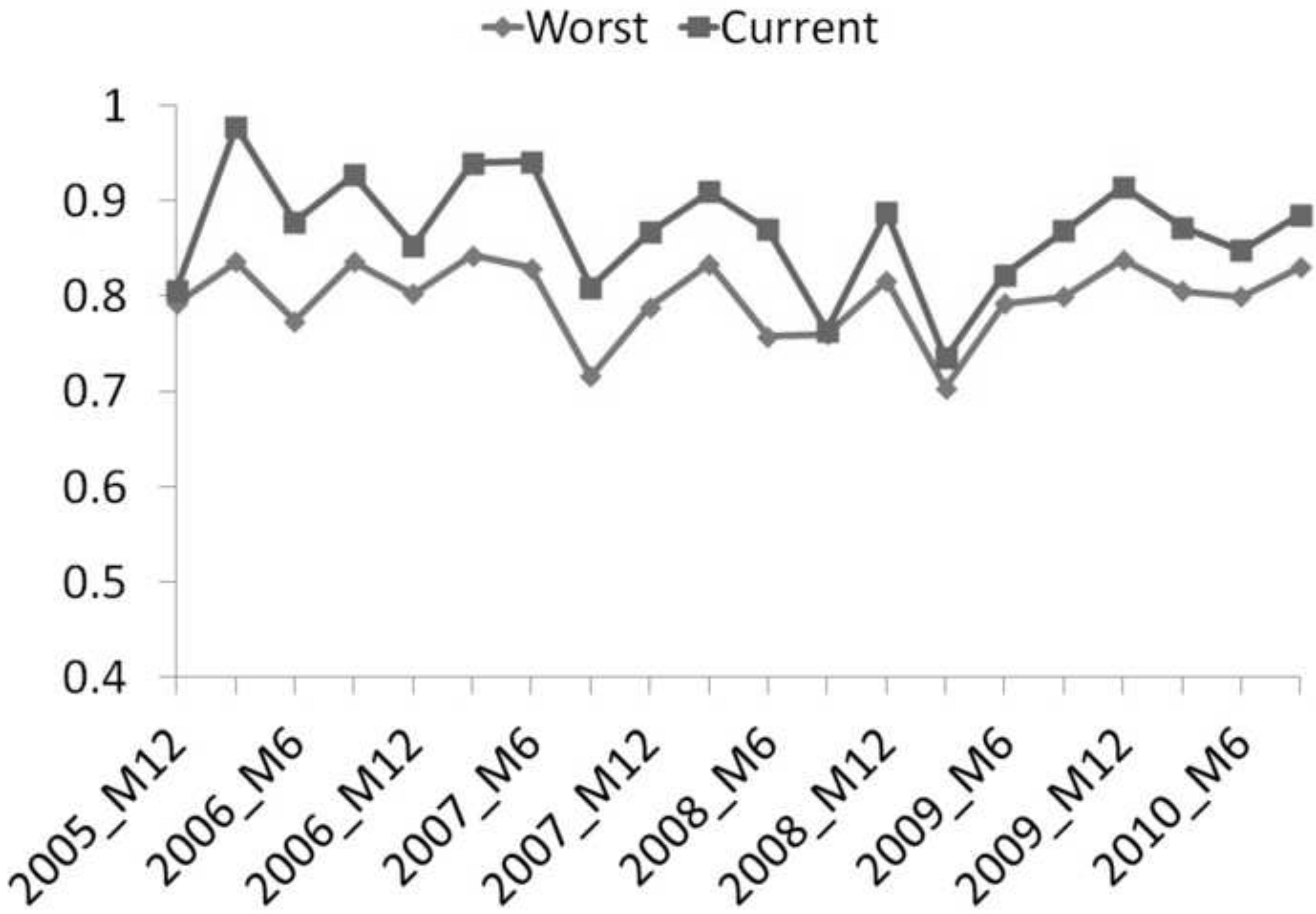


Figure 6b greyscale

Click here to download high resolution image

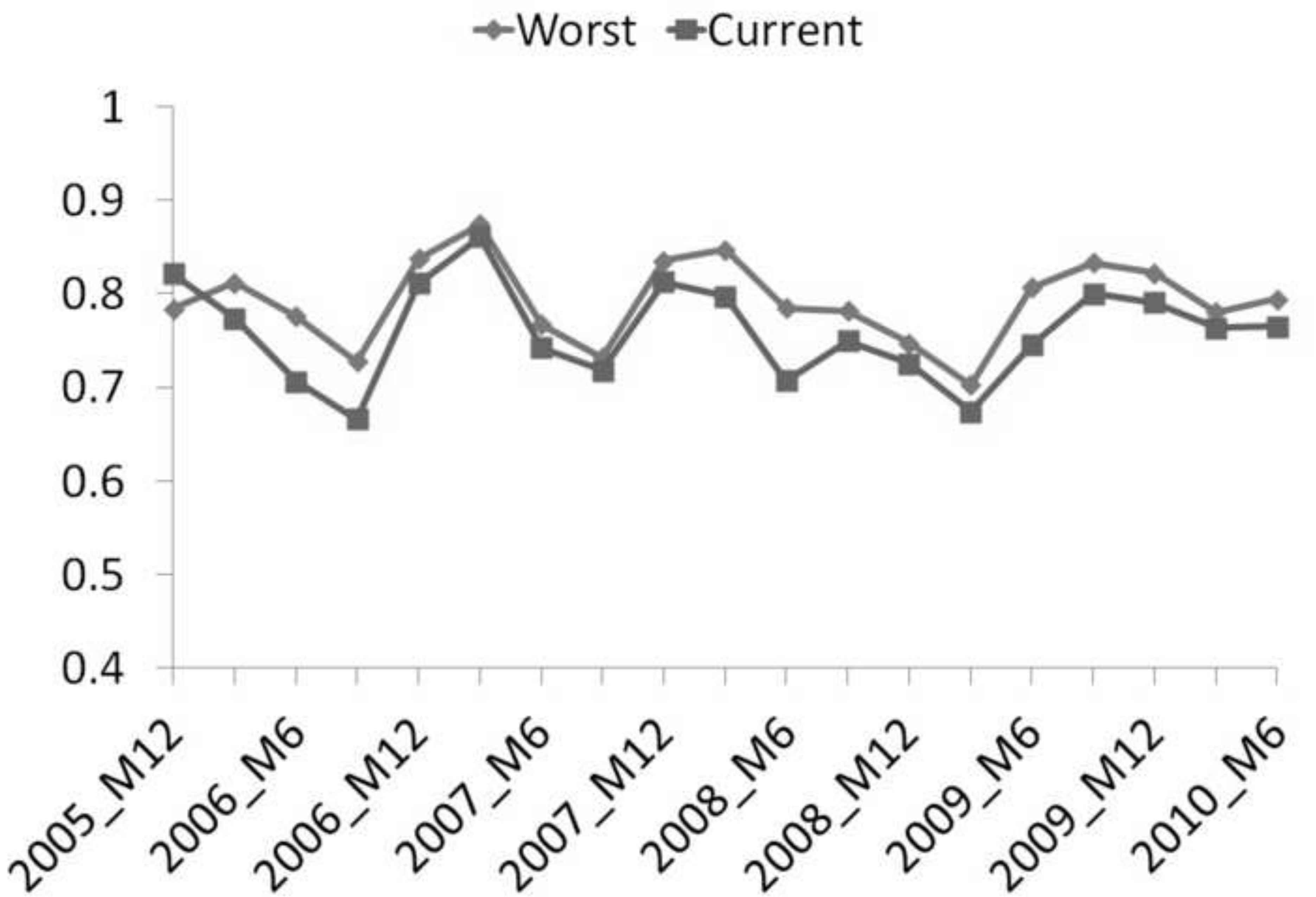


Figure 6c greyscale

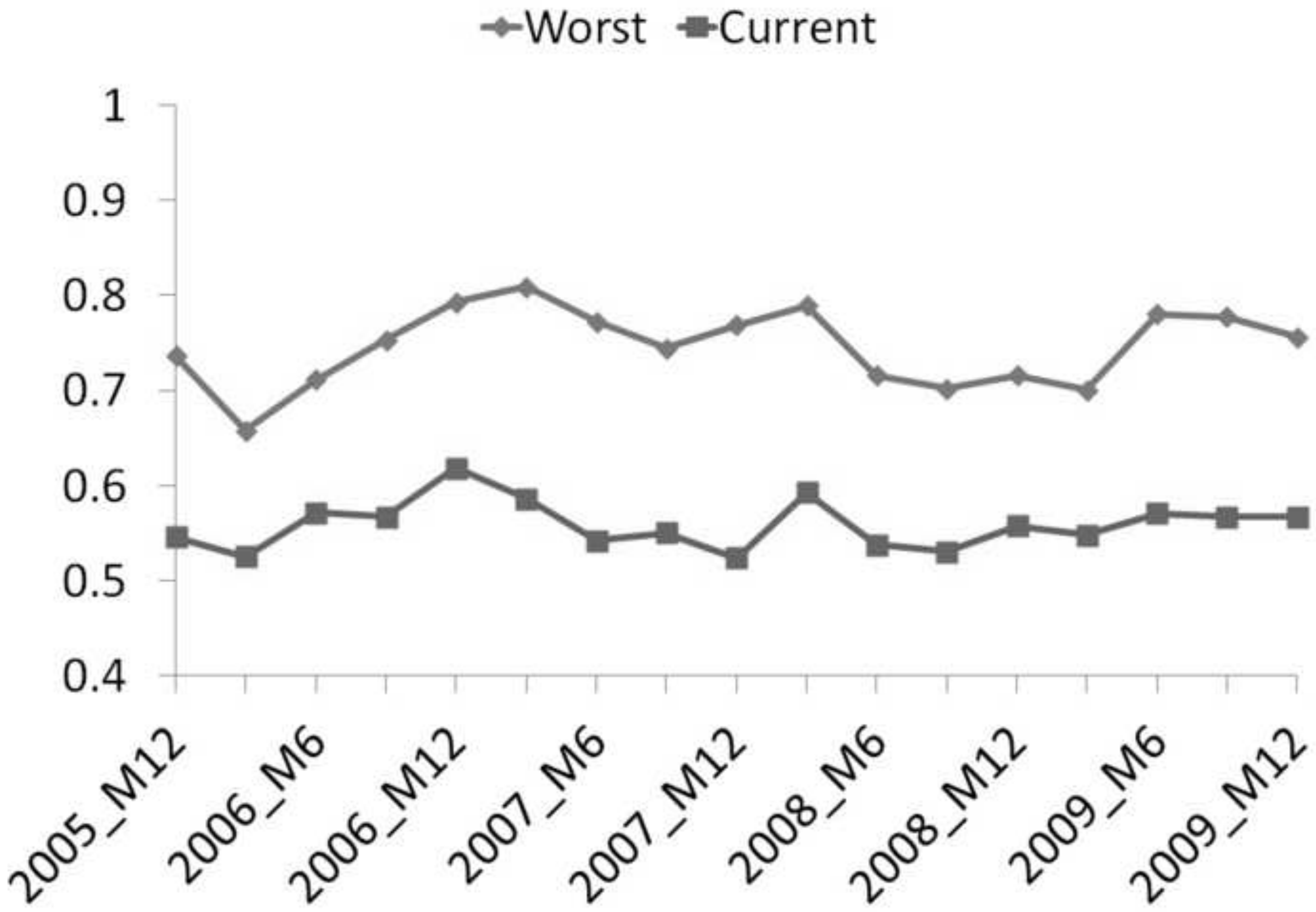
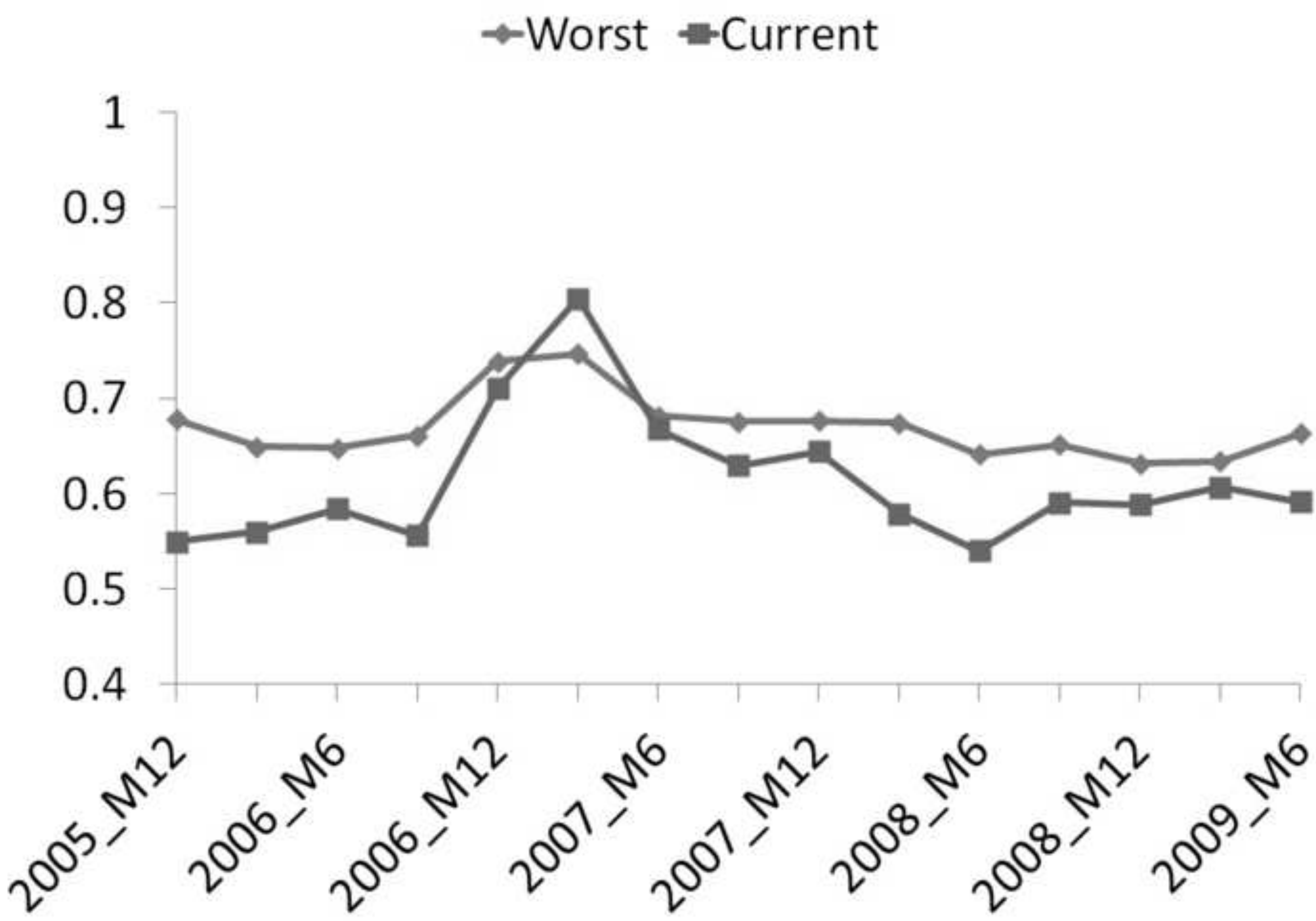
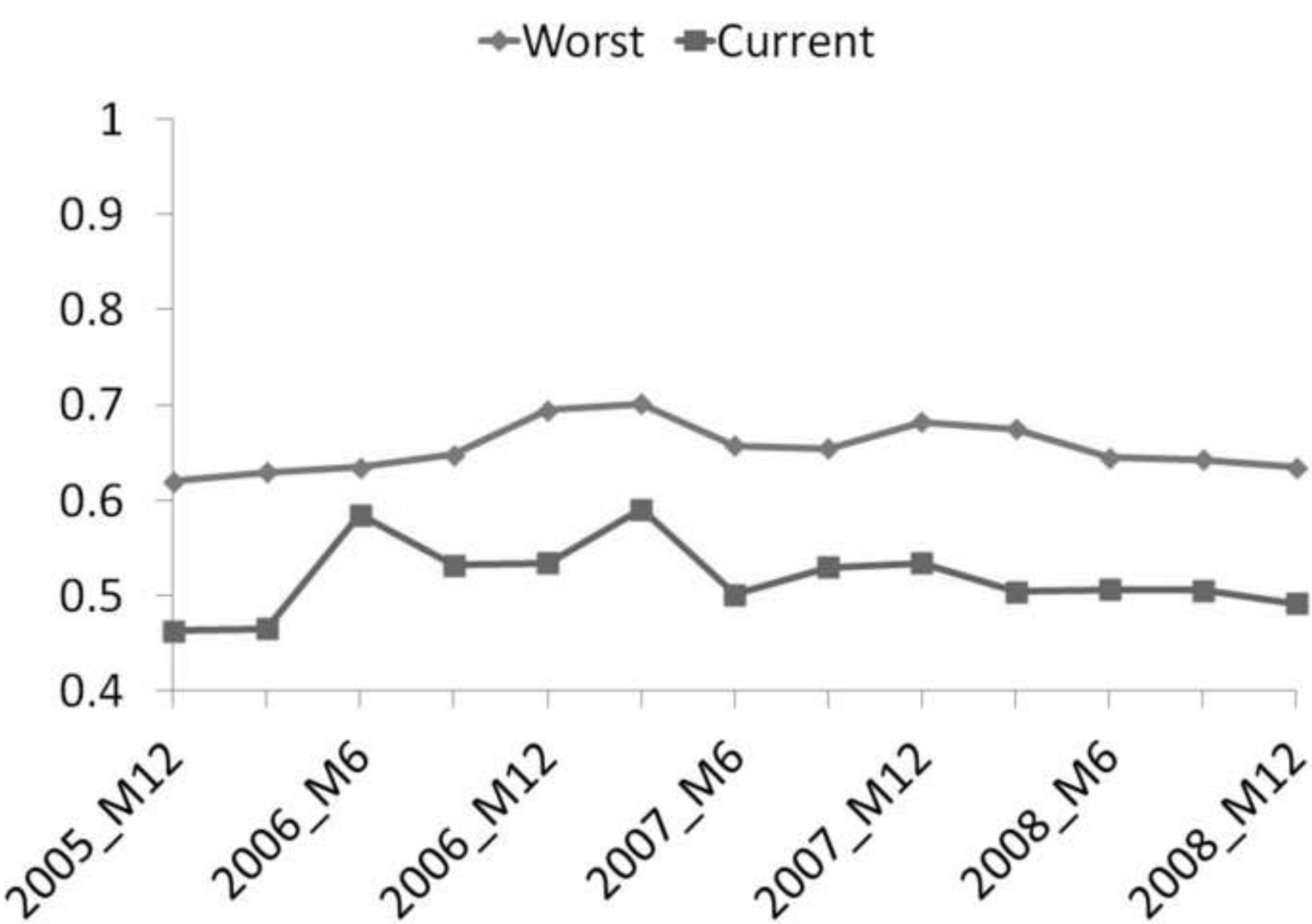


Figure 6d greyscale



Figure 6e greyscale
Click here to download high resolution image