INSTITUT DES ACTUAIRES

100% DATA SCIENCE

NON-PARAMETRIC INDIVIDUAL CLAIM RESERVING IN INSURANCE

MAXIMILIEN BAUDRY* & CHRISTIAN ROBERT*

CHAIR OF EXCELLENCE
Data Analytics & Models for Insurance

*DAMI Chair & Université Lyon 1 - ISFA

http://chaire-dami.fr/

ACKNOWLEDGEMENTS TO

- The actuarial department of BNP PARIBAS CARDIF
- And its Data Lab datalab'
- And more specifically to Philippe Baudier, Pierre de Sahb and Sebastien Conort

RESERVING 2.0: CHANGING THE WAY OF THINKING ABOUT OUTSTANDING CLAIMS

- Computing reserves by leveraging the use of massive data and artificial intelligence in insurance.
- Taking advantage of Machine Learning algorithms' predictive power.
- Exploiting the richness of the insurer's data (**any** sort of structured and unstructured data).
- Deploying this method on **any** kind of portfolio (life, non-life).

INTRODUCTION & MOTIVATIONS

- The **current reserving practice** consists, in most cases, in using methods based on **claim development triangles**.
- Triangles are organized by **origin period** (occurrence most of the time or underwriting otherwise) **and development period**.
- **Deterministic and stochastic unpaid claim reserving models** based on triangles (e.g. Chain Ladder method, Bornhuetter-Ferguson method) have had a great success to **manage reserve risk** for a variety of lines of business...

- ... but such models suffer form underlying strong assumptions and give rise to **several issues** :
  - Need for tail factors that may induce **over parameterization risk**.
  - **Propagations of errors** through the development factors, huge estimation error for the latest development periods.
  - **Instability in ultimate claims for recent arrival years**, uncertainty about the ability to properly capture the pattern of claim development.
  - Lack of robustness and need for **treatments of outliers**.
  - **Can not separate assessment of IBNR and RBNS claims**.
  - ...

STATE OF THE ART

- Because triangle-based methods use **aggregated data**, they **don't use any information on the policy, the claim nor the policy holder**.

- Natural overcome: **Individual claim reserving**.
  - First approaches: **Structural and Parametric models**.
  - Recent approaches (very few): **Non-parametric & Machine Learning models**. (Our approach).

WORKS ON MACHINE LEARNING (ML) MODELS IN RESERVING

- **Wüthrich (2017)**
  - Prediction on the number of RBNS payments with ML (CART algorithm).
  - IBNR prediction with Chain Ladder (CL) method.
- **ASTIN ICDML working group (2017)**
  - No use of explanatory variables/features (only payments).
  - 'Cascade' predictions, i.e. chaining predictions, inducing propagation of errors just like Chain Ladder does.
  - No IBNR prediction at all. The reporting delay is forced to 0.
  - Questionable data simulation: each payment (cumulative) is the product of the ultimate with a noisy coefficient.

WHAT WE PROPOSE

- A new non-parametric and flexible approach to estimate individual claims reserves which handles key effects, such as:
  - Including the key claim characteristics (i.e., explanatory variables) to **allow for claims heterogeneity** and to take advantage of additional large datasets.
  - Learning the specific development pattern of claims, including their occurrence, reporting and cash-flow features, and **detecting potential trend changes**.
  - Taking into account possible deviations in the product mix, the legal context or the claims processing over time, to **avoid potential biases in estimation and forecasting**.
  - Implementing **separate** and **consistent treatments of IBNR and RBNS claims**.

- Our model is estimated on **simulated data** and the prediction results are compared with those generated by the Chain Ladder model.

- When evaluating the performance of our approach, we put emphasis on the the **impact of using micro-level information on the variances of the prediction errors**.

- We implement our new approach with an ExtraTrees algorithm but **many other powerful machine learning algorithms can easily be adapted (random forest, gradient boosting,...)**.

MODULARITY – PREDICTION FLOW

- **Our prediction flow is able to**:
  - Compute **RBNS and IBNR reserves separately**.
  - Run the full reserving process as of **any date** as required (backtesting…).
  - Compute reserves along **any time granularity (monthly, quarterly, semestrial, annual)**.
  - Learn from **any subsample of historical data** specified by user.
  - Learn from **any sub-space of features** specified by the user.

THE PROBLEM & OUR APPROACH

- We associate with **each policy** the following quantities :

- T_0 : the **underwriting date** (Δ is the **insured period** and the contract will expire at $T_0 + \Delta$).

- Some **features/risk factors** are known at T_0 and may evolve over time : $(F_t)_{t \geq T_0}$

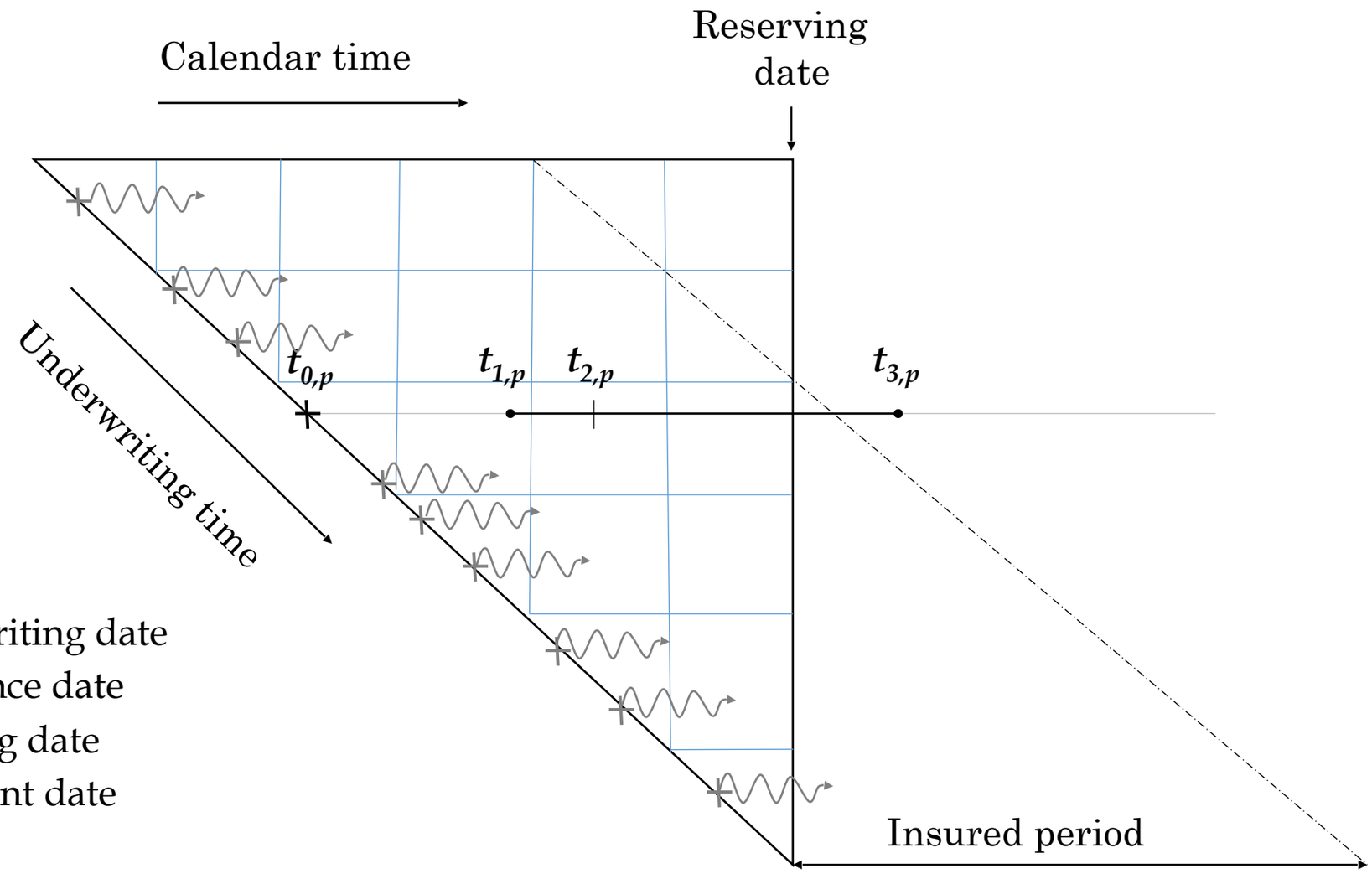
Example: For a life insurance policy : applicant's current age, applicant's gender (if allowed), height and weight of the applicant, health history, applicant's marital status, applicant's children, if any..., applicant's occupation, applicant's income, applicant's smoking habits or tobacco use)...

- T_1 : the **occurrence date** of the claim ($T_1 = \infty$ if there is no claim). Only one claim is possible during the insured period (but it can be easily generalized).

- T_2: the **reporting date**.
  - We assume that there exists a **maximum delay** $\Delta_{\max,r}$ to report the claims once it has occurred, i.e. $T_2 - T_1 < \Delta_{\max,r}$
- T_3: the **settlement date**.
  - During the **settlement period** the insurance company receive **information on the individual claim** like exact cause of accident, type of accident, location of accident, line-of-business and contracts involved, claims assessment and predictions by claims adjusters, payments already done, external expertise, etc.
  - We denote this information by $(I_t)_{t \geq T_2}$
  - We assume that there exists a **maximum delay** $\Delta_{\max,s}$ to settle the claims once it has been declared, i.e. $T_3 - T_2 < \Delta_{\max,s}$

- **Payment cash flows**
  - The payments are broken down into q several components : $q-1$ insurance coverages and the legal and claims expert fees (if any).
  - We denote by $(P_t)_{T_2<t\leq T_3}$ the **cumulated payment process**.
    We let $P_t = 0$ for $T_1 < t \leq T_2$

- **External information** may be used to predict reserves. We denote by E_t that information.



Policy p :

$t_{0,p}$ underwriting date

$t_{1,p}$ occurrence date

$t_{2,p}$ reporting date

$t_{3,p}$ settlement date

CATEGORIES OF OUTSTANDING CLAIMS

Note that if $T_1 > T_0 + \Delta$, the insurance company is not liable for this particular claim with the actual insurance policy because the contract is already terminated at claim occurrence.

1. $t < T_1$. **There is no outstanding claim**.
2. $T_1 < t < T_2$, **The insurance claim has occurred but it has not yet been reported to the insurance company**.

These claims are called **Incurred But Not Reported (IBNR)** claims. For such claims:

- We do not have individual claim specific information.
- But we can use any external information E_t

$$IBNR_t = \mathbb{E}\left[P_{T_3} 1_{T_1 < t \wedge (T_0+\Delta)} | t < T_2, (F_u)_{T_0 \leq u \leq t}, (E_u)_{0 \leq u \leq t}\right]$$

3. $T_2 < t < T_3$. **These claims are reported at the company but the final assessment is still missing**.

Typically, we are in the situation where more and more information about the individual claim arrives, and the prediction uncertainty in the final assessment decreases.

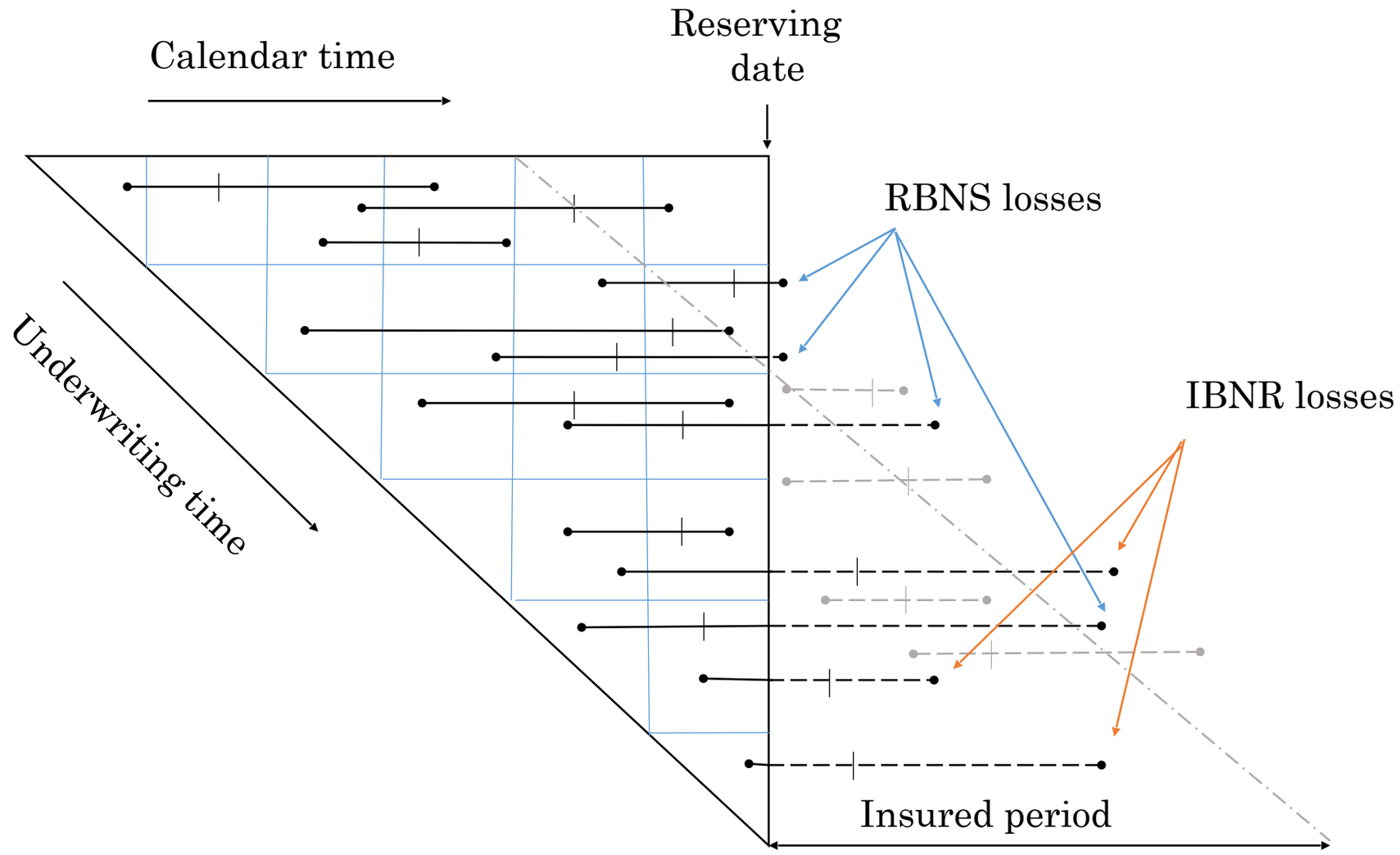
However, these claims are not completely settled, yet, and therefore they are called **Reported But Not Settled (RBNS)** claims :

$$RBNS_t = \mathbb{E}\left[P_{T_3} - P_t | T_2 < t < T_3, T_1 < T_0 + \Delta, (F_u)_{T_0 \leq u \leq t}, (E_u)_{0 \leq u \leq t}, (I_u)_{T_2 \leq u \leq t}\right]$$

$\Rightarrow$ **The individual claims reserve is therefore:**

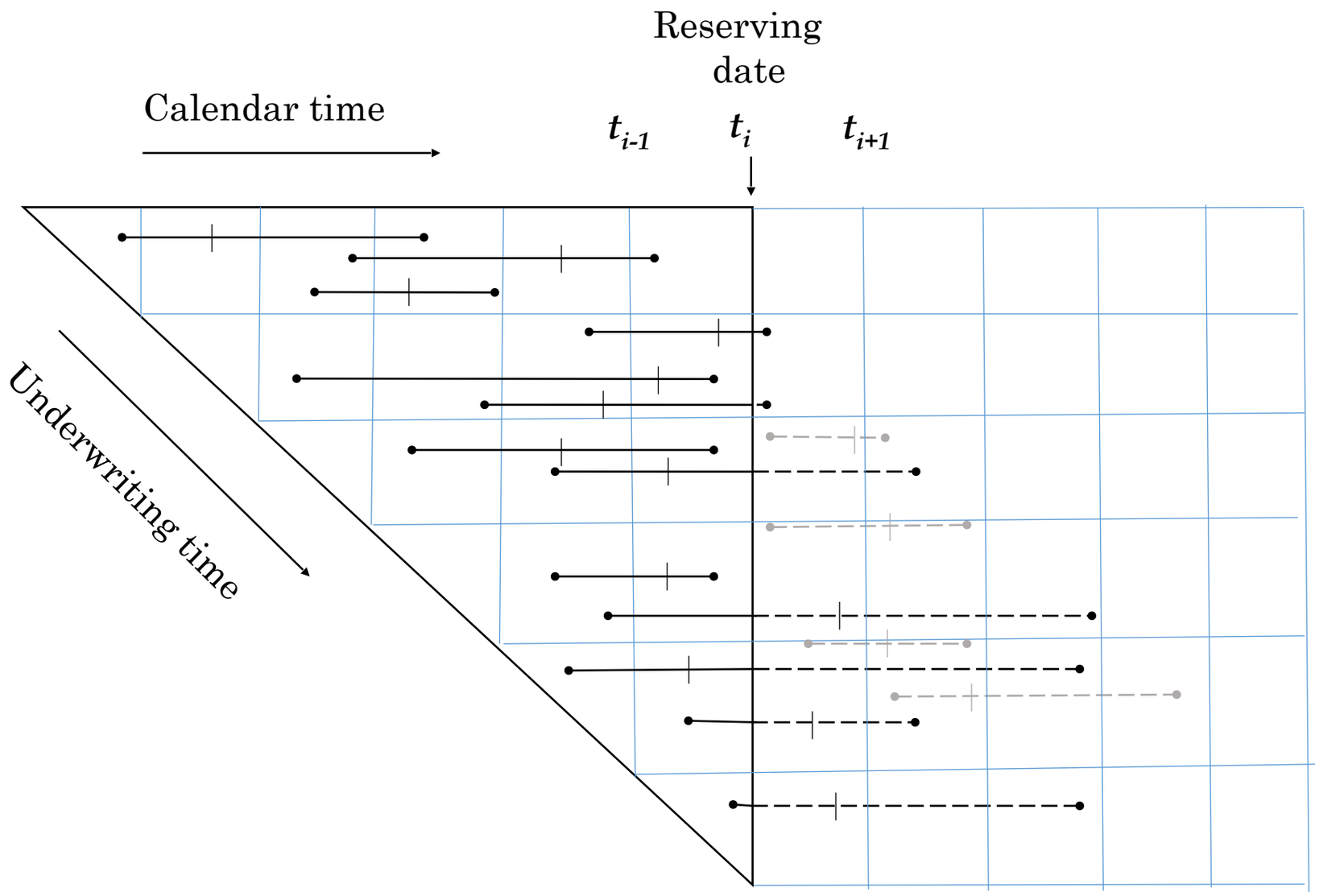
$$ICR_t = IBNR_t 1_{t<T_2} + RBNS_t 1_{t \geq T_2}$$

SCHEME OF OUR FRAMEWORK



SUBDIVISION OF OUTSTANDING CLAIMS

- Let δ be a fixed timestep and derive a **grid of times** $t_i = \delta \times i,\ i \geq 0$, for which the **insurance company wants to evaluate its liabilities**.



- We **split** $RBNS_{t_i}$ in the following way : for $j = 1, 2, 3, ...$ we define the expected increase of the payments between t_{i+j-1} and t_{i+j} given that a claim has been declared

$$RBNS_{t_i,j} = \mathbb{E}\left[P_{t_{i+j}} - P_{t_{i+j-1}} | T_2 < t_i < T_3, T_1 < T_0 + \Delta, (F_u)_{T_0 \leq u \leq t_i}, (E_u)_{u \leq t_i}, (I_u)_{T_2 \leq u \leq t_i}\right]$$

such that

$$RBNS_{t_i} = \sum_{\text{claim development period } j} RBNS_{t_i,j}.$$

- We **split** $IBNR_{t_i}$ in the following way : for $j = 1, 2, 3, ...$ we define the expected increase of the payments between t_{i+j-1} and t_{i+j} given that a claim has been declared

$$IBNR_{t_i,j} = \mathbb{E}\left[(P_{t_{i+j}} - P_{t_{i+j-1}})1_{T_1<t_i\wedge(T_0+\Delta)}|t_i < T_2, (F_u)_{T_0\leq u\leq t_i}, (E_u)_{T_0\leq u\leq t_i}\right]$$

such that

$$IBNR_{t_i} = \sum_{\text{claim development period } j} IBNR_{t_i,j}.$$

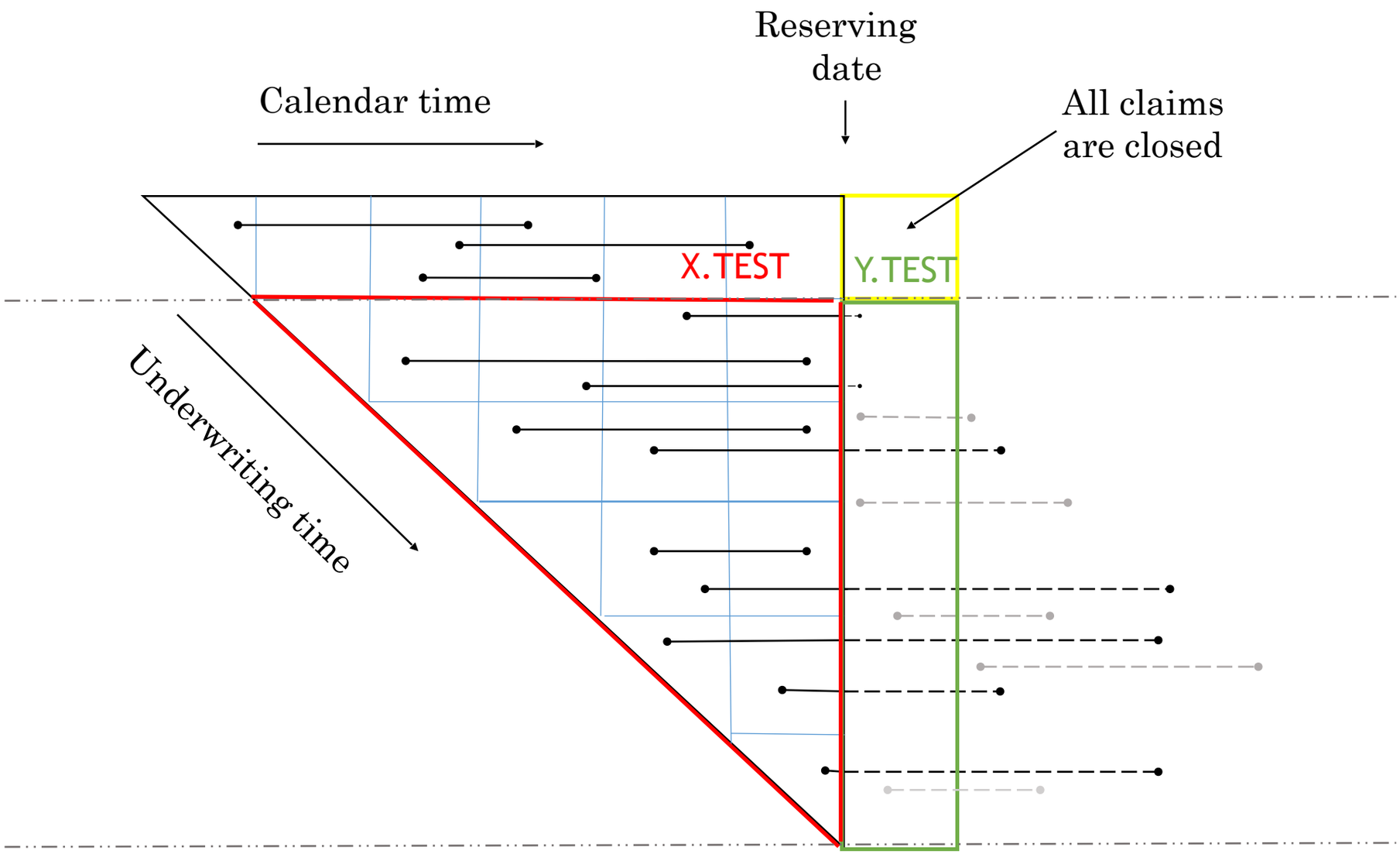
Moreover, we write $IBNR_{t_i,j}$ in a **frequency/severity formula**:

$$IBNR_{t_i,j} := IBNR\_freq_{t_i,j} \times IBNR\_loss_{t_i,j}$$

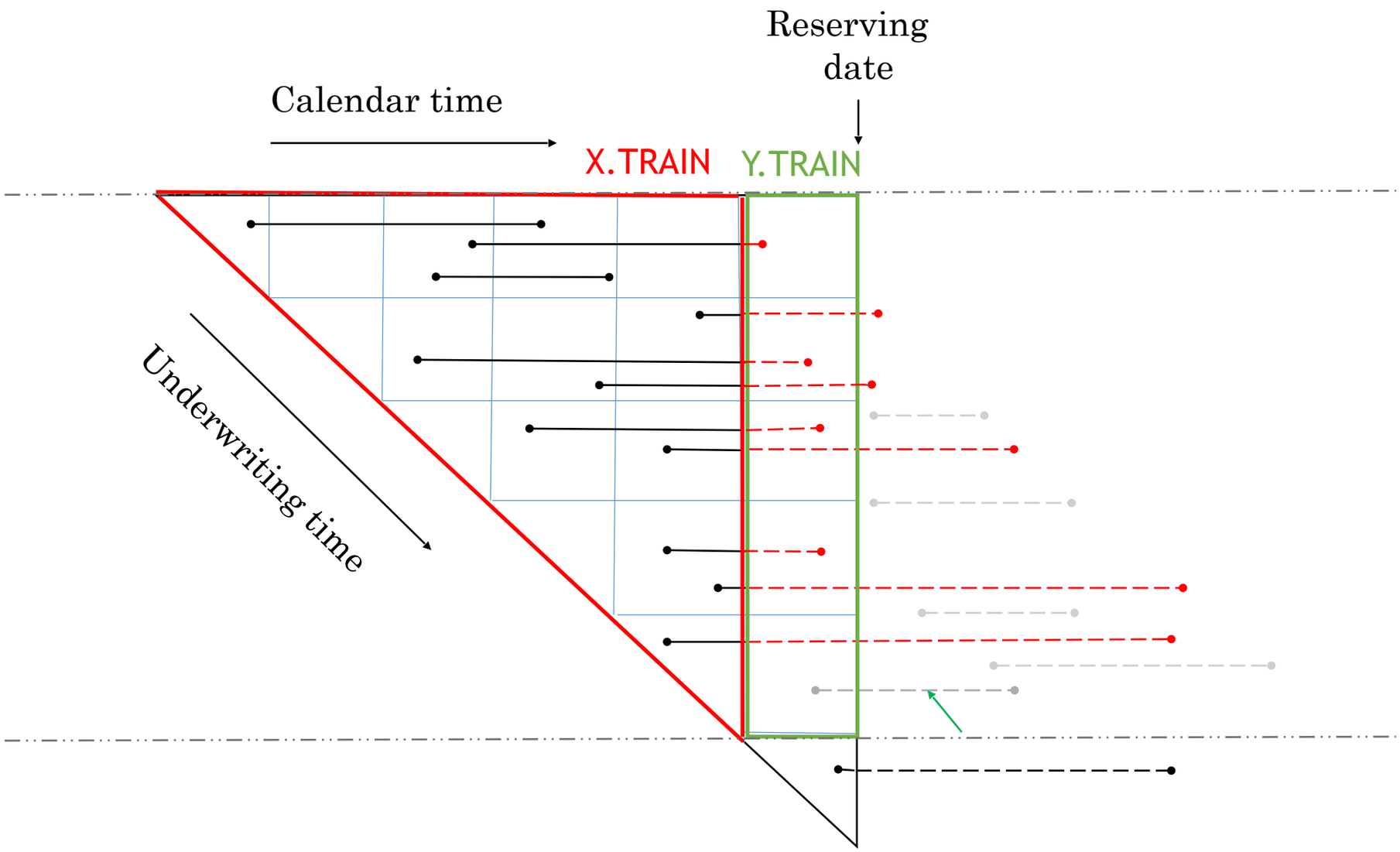
DATABASE BUILDING FOR THE MACHINE LEARNING APPROACH AND PREDICTIONS

- Notations:
  - Reserving date: t_i
  - Development period: j
  - Model: k

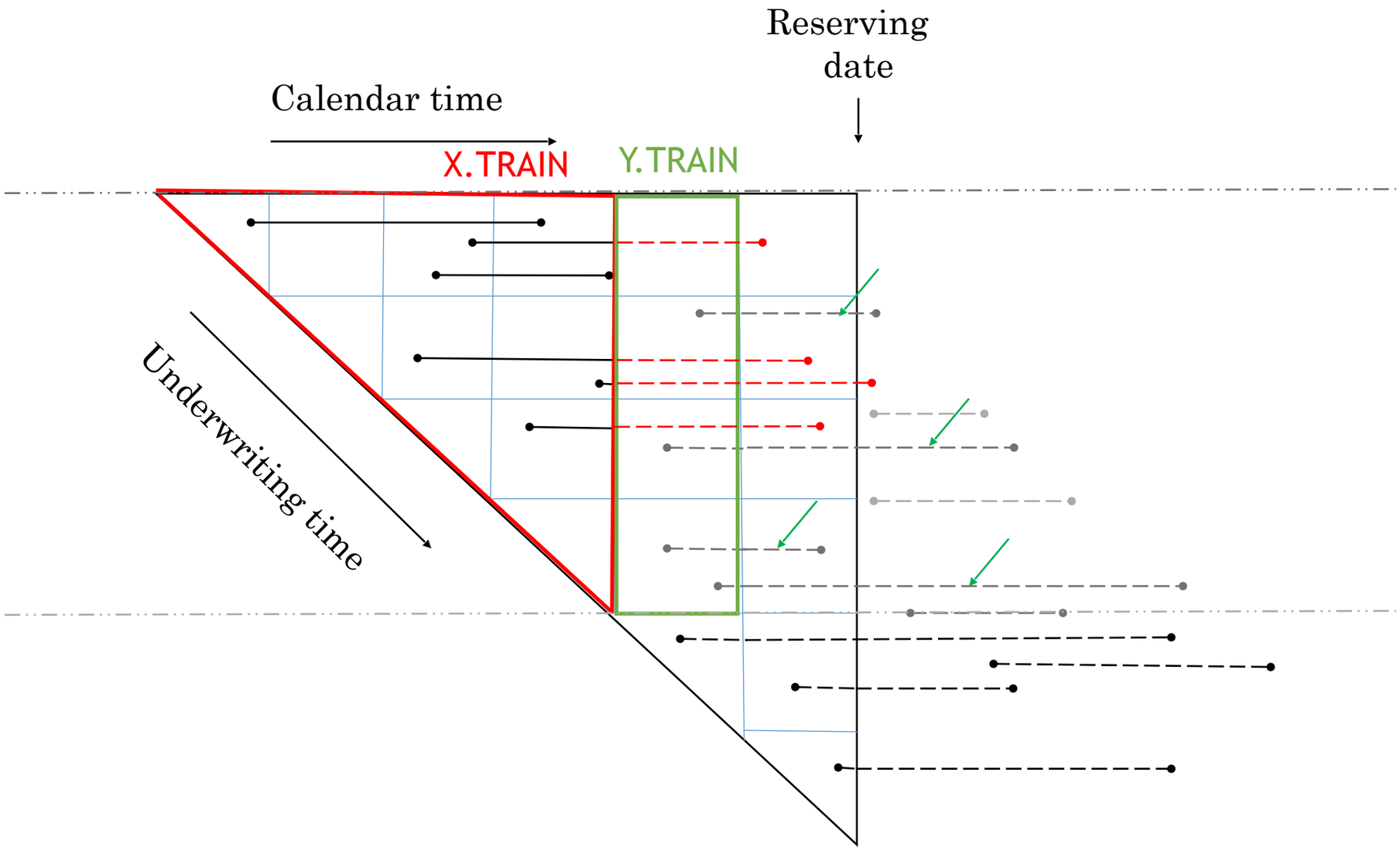
- **Case** $j = 1$: 1st development period - Test sets



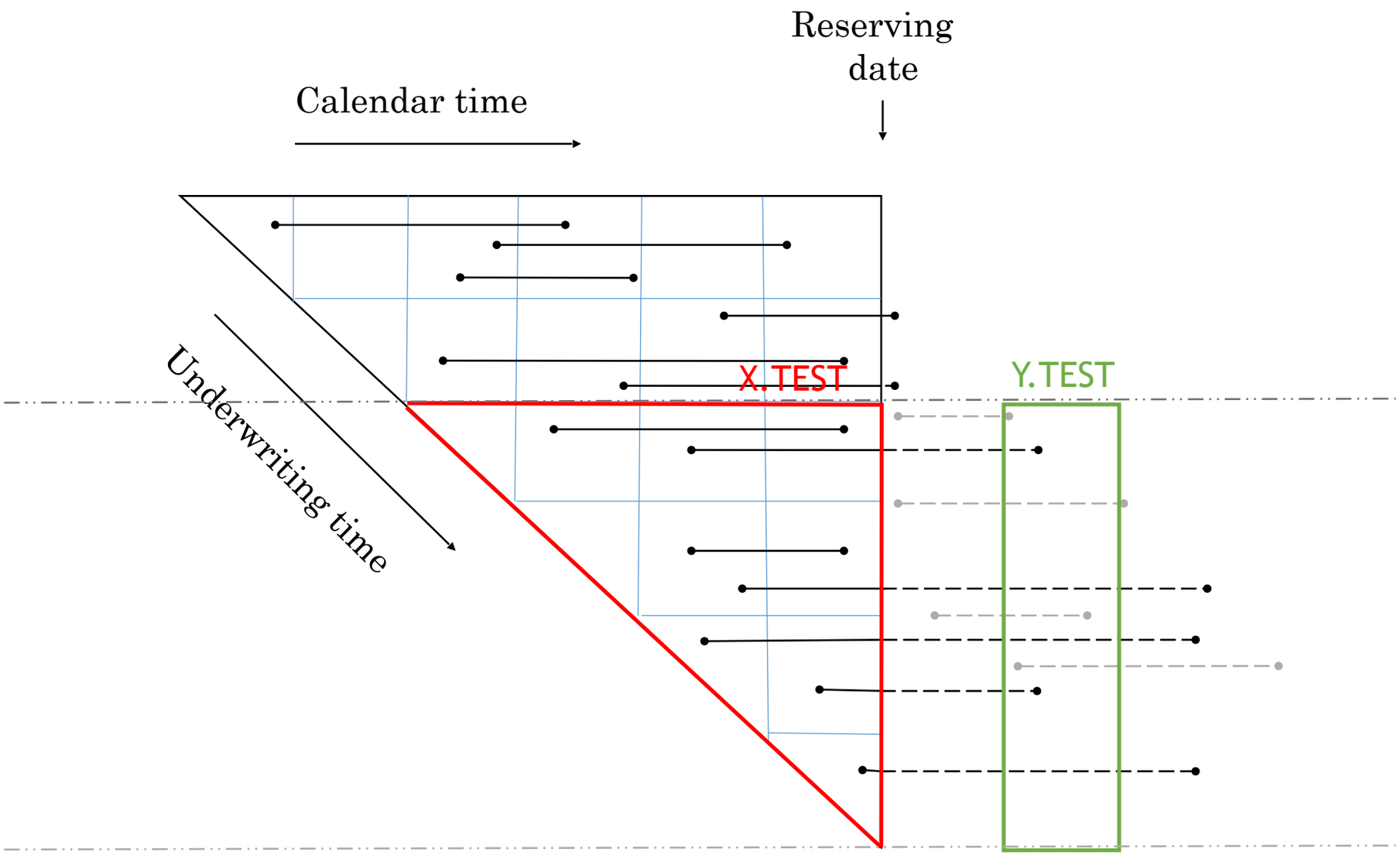
- **Case** $j = 1,\ k = 1$: 1st development period - Train sets - 1st model



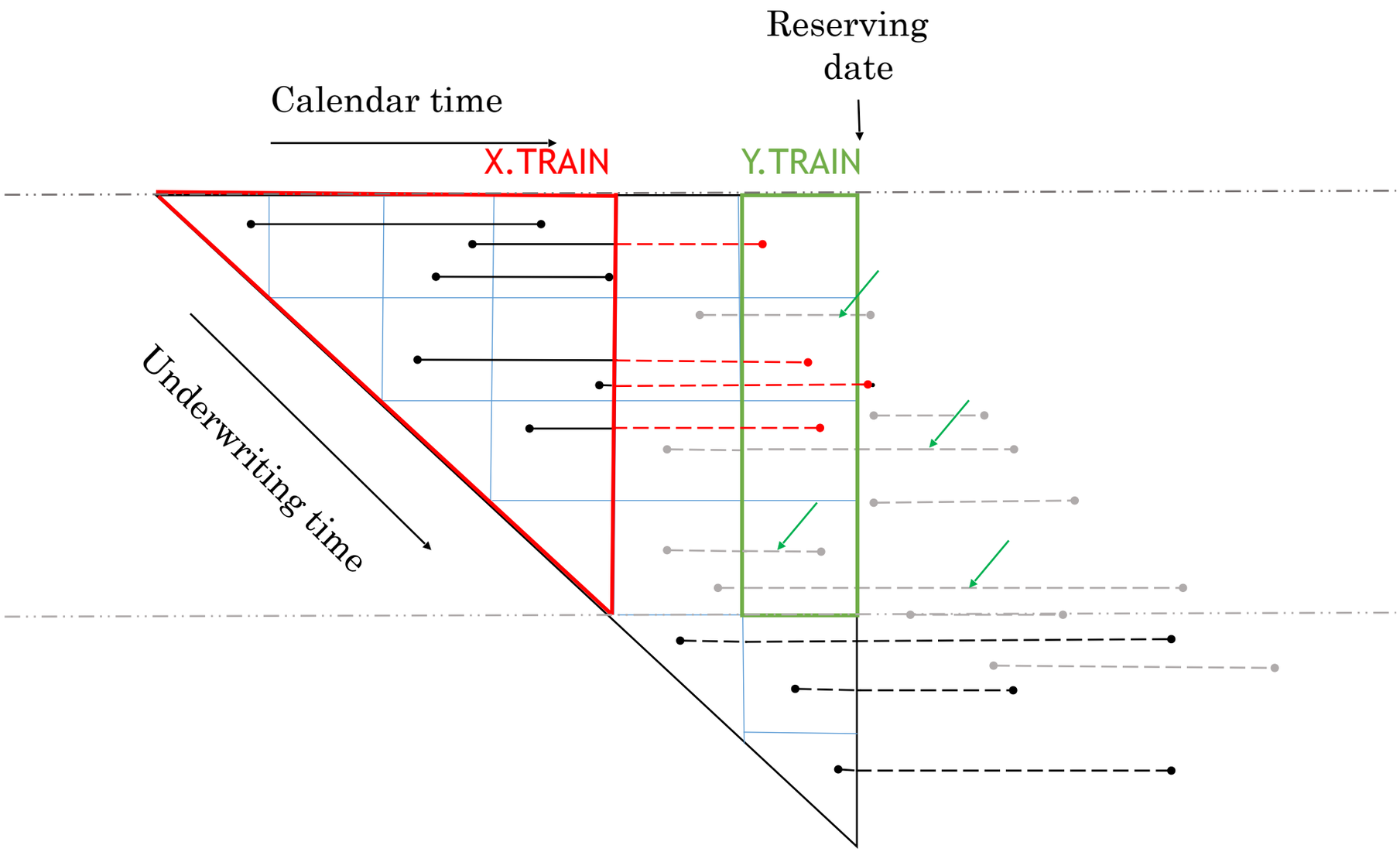
- **Case** $j = 1,\ k = 2$: 1st development period - Train sets - 2nd model



- **Case** $j = 2$: 2nd development period - Test sets



- **Case** $j = 2,\ k = 1$: 2nd development period - Train sets - 1st model



- **Final individual claims reserve predictions:**

$$\widehat{ICR}_{t_i,p} = \widehat{IBNR}_{t_i,p} 1_{t_i < T_{2,p}} + \widehat{RBNS}_{t_i,p} 1_{t_i \geq T_{2,p}}$$

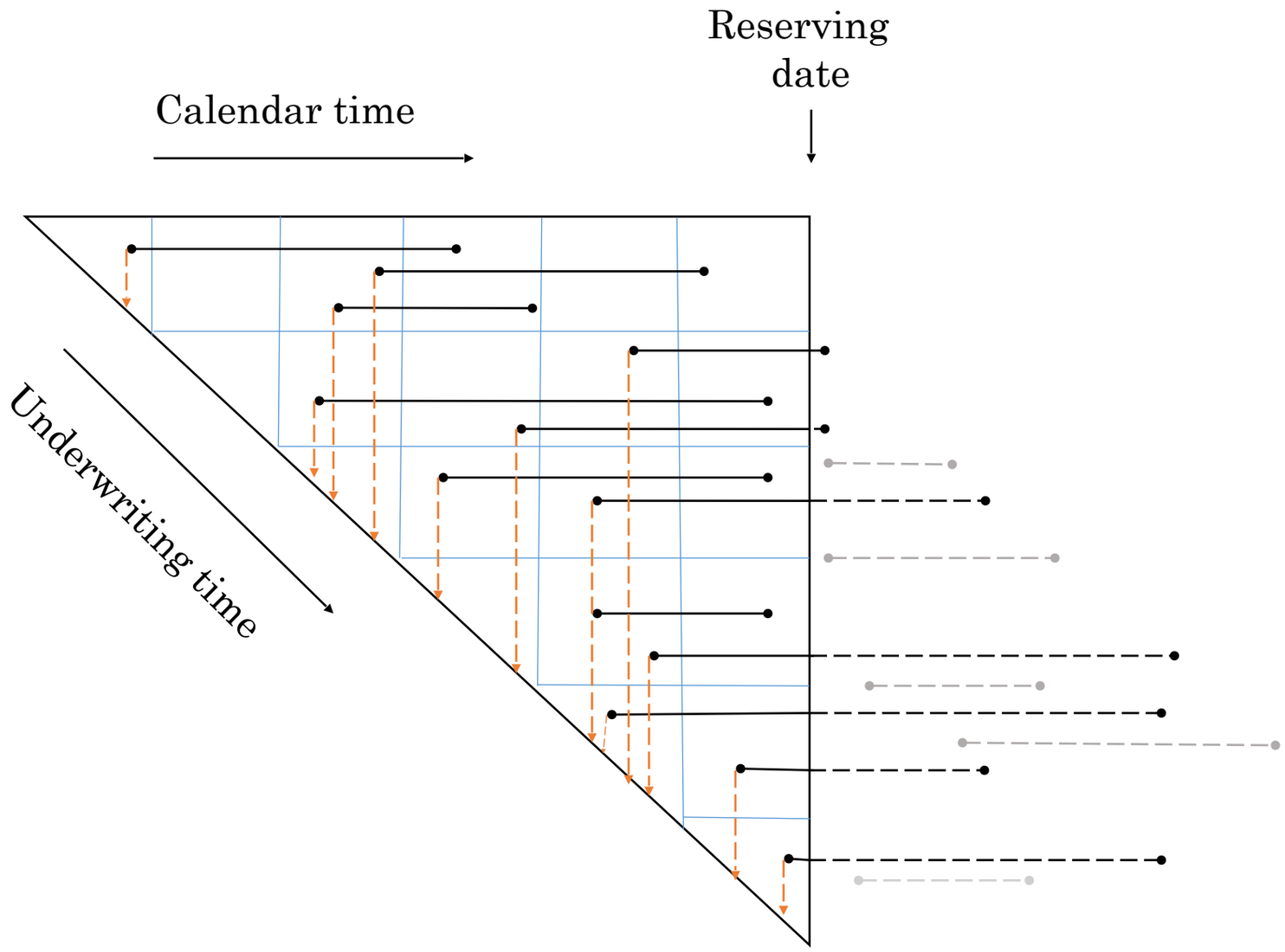
- **Final claims reserve prediction:**

$$\sum_{p \in RBNS \cup IBNR} \widehat{ICR}_{t_i,p}$$

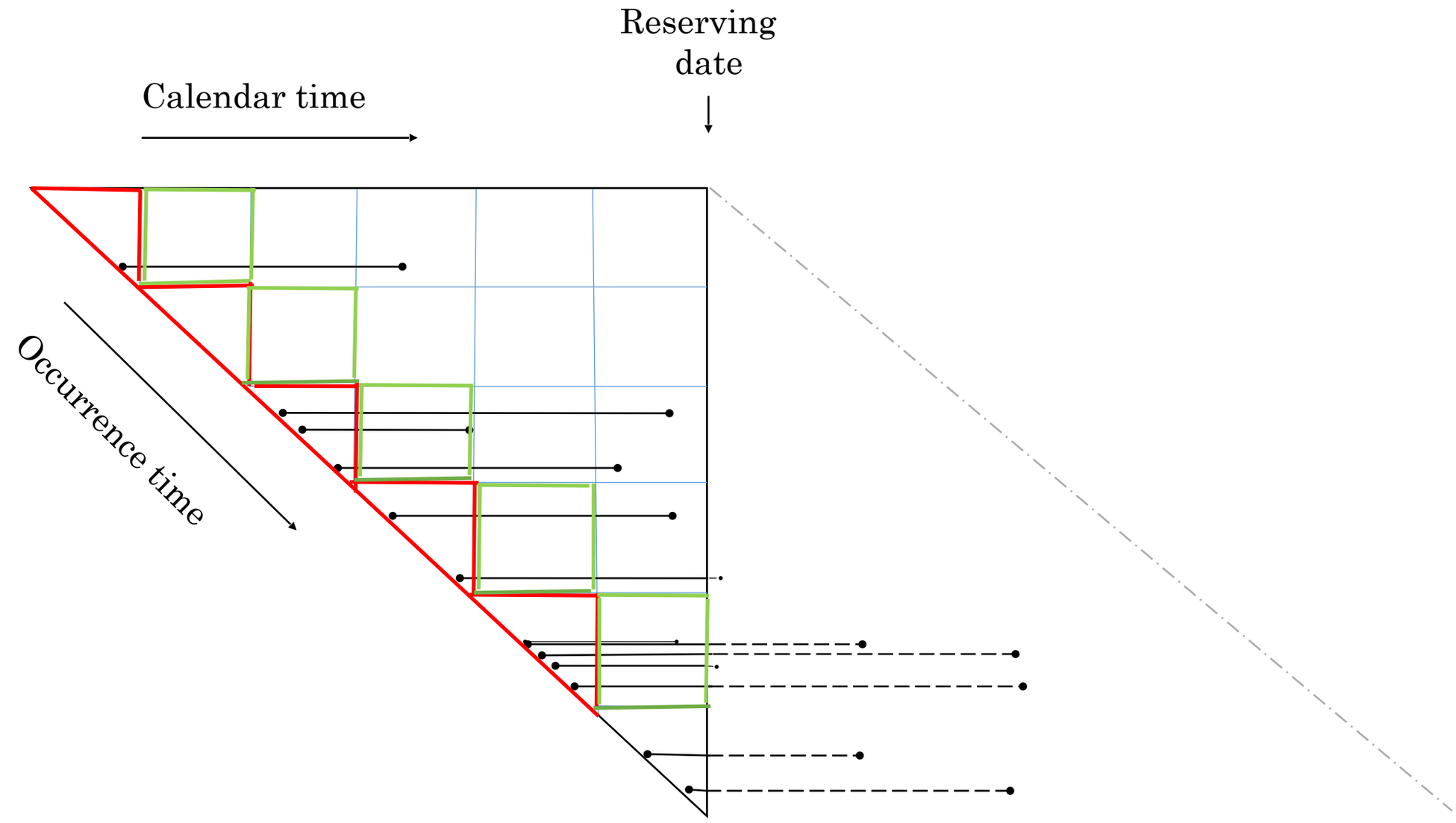
DATABASE BUILDING FOR THE CHAIN LADDER APPROACH AND PREDICTIONS

- **Notations:**
  - Reserving date: t_i
  - Development period: j

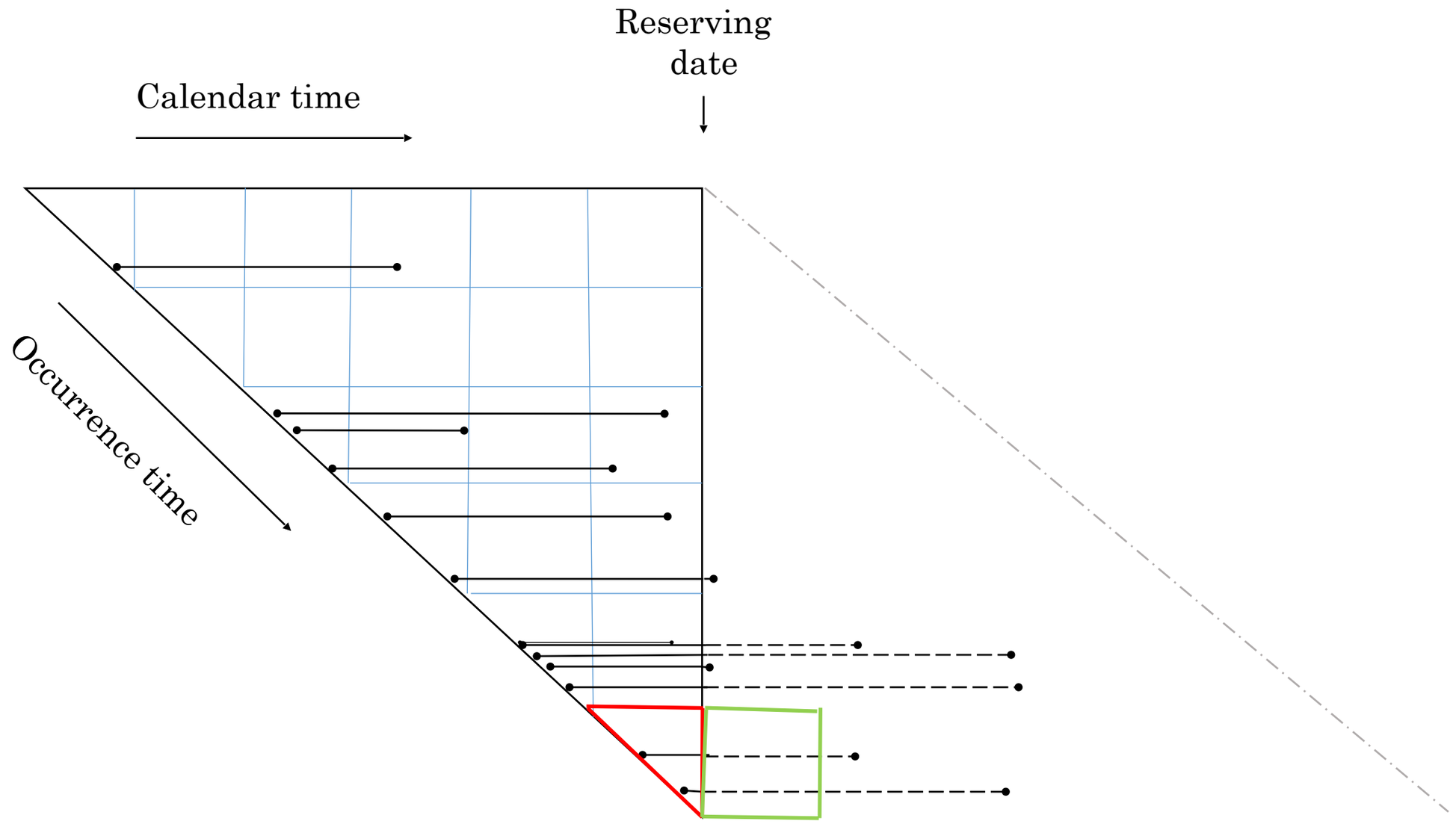
- **From underwriting time to occurrence time:**



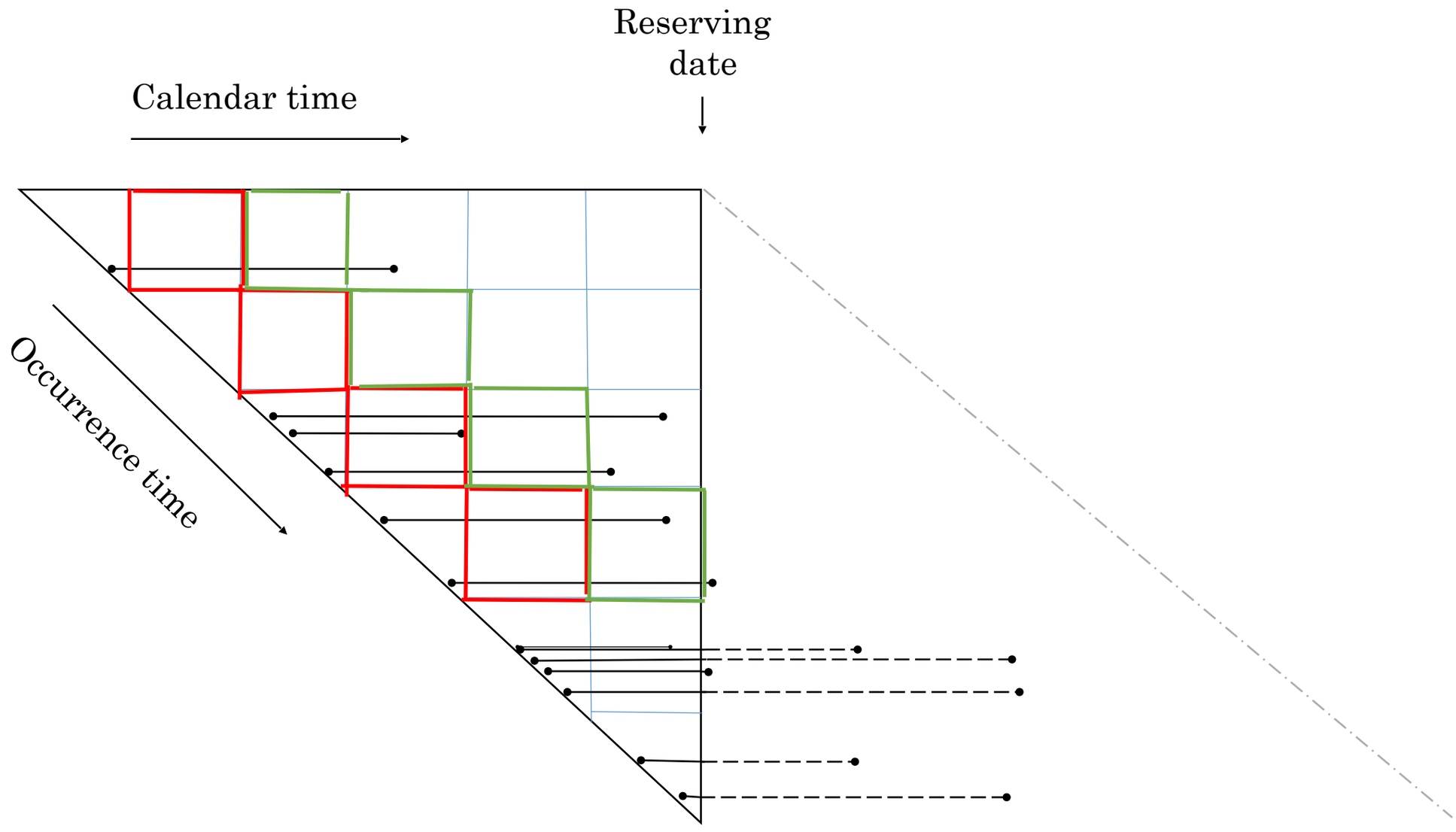
- **Example:** 1st development period



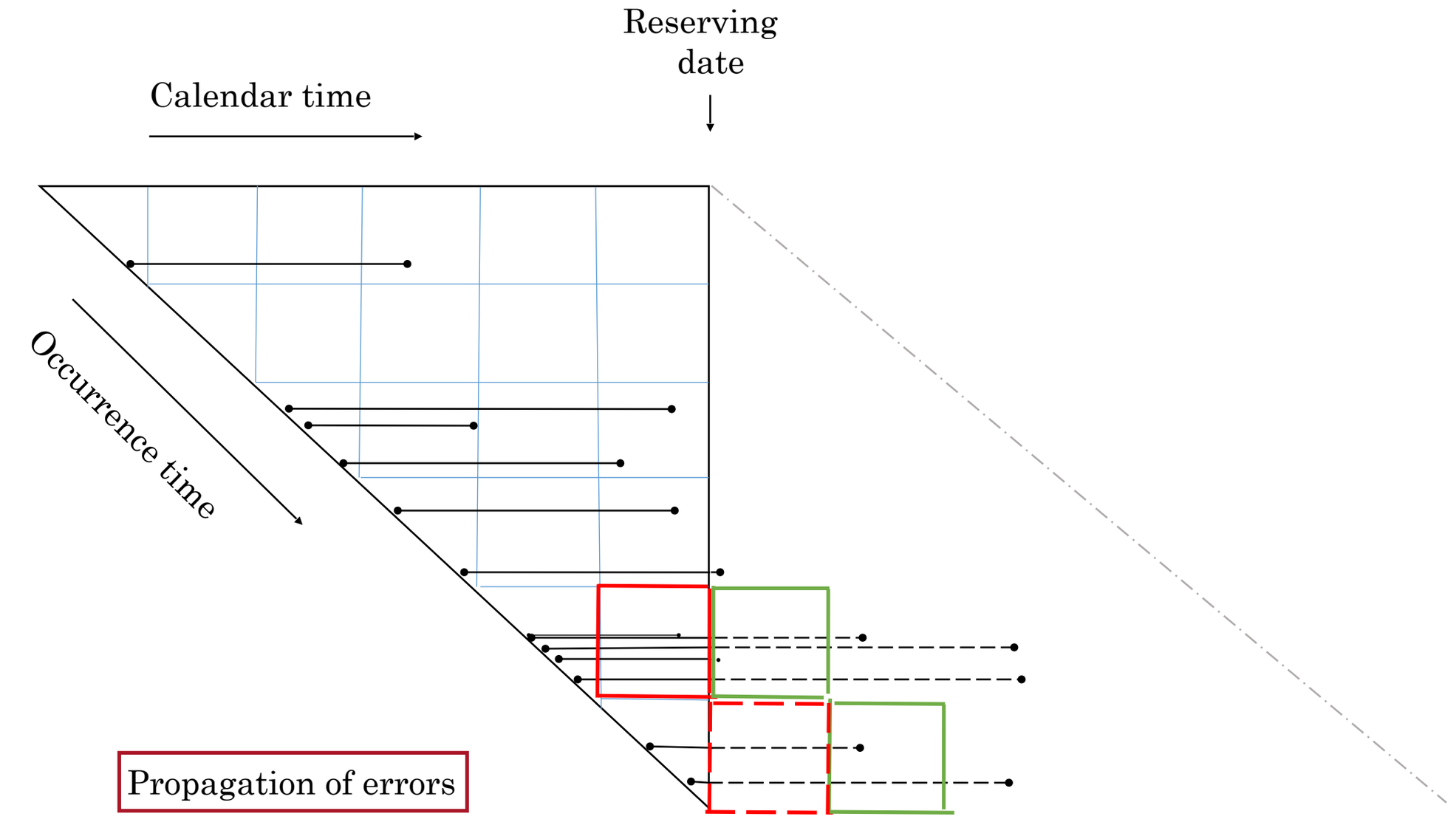
- **Example:** 1st development period prediction



- **Example:** 2nd development period



- **Example:** 2nd development period prediction



PREDICTION ALGORITHM

- The Extra-Trees algorithm builds an **ensemble of unpruned regression trees** according to the classical top-down procedure.
- Its two main differences with other tree-based ensemble methods are that:
  - it splits nodes by choosing cut-points fully at random.
  - it uses the whole learning sample (rather than a bootstrap replica) to grow the trees.
- The predictions of the trees are aggregated to yield the final prediction, by majority vote in classification problems and arithmetic average in regression problems.

Remarks:

1. The choice of the prediction algorithm is not the main issue here as long as we use a non-parametric algorithm such as RandomForest, ExtraTree, XGBoost etc…
2. We could use a GLM model, but this needs work on interaction detection, feature engineering and additional assumptions (such as the distribution and link function) which must be tested.

A FEW NUMBERS ON PROGRAMMING

- We used Python to develop our method.
- It runs in 2 steps:
  - Build train and test sets.
  - Predict payment increases for each claim development period.
- ~2k rows of code.
- On the following example, runs in a few minutes for a relatively big portfolio on a huge hardware (for one seed):
  - 80 vCPU.
  - 128 Go RAM.

A CASE STUDY WITH MOBILE PHONE INSURANCE

- **Note that every characteristic in our portfolio is customizable. This allows us to work with very different kind of portfolio.**
- We consider a **mobile phone insurance** which covers the devices for the following damages:
  - **Theft**
  - **Breakage**
  - **Oxidation**
- The insurance company provides cover for a range of four phone brands and up to four models by brand with three policy types available for an **insured period of one year:**
  - **“Breakage”**
  - **“Breakage and oxidation”**
  - **“Breakage, oxidation and theft”**

CENTRAL SCENARIO

- For the first generation of policies which will be sold from 2016/01/01 to 2017/12/31, we consider the following central scenario :

  - The underwriting Poisson point process has a constant intensity $\lambda_{0,t} = 250,000$ (in yearly unit), i.e. the insurance sells roughly $500,000$ policies over the two years.

  - Claim severity distributions: Beta distributions

Damage type	α	β
Breakage	2	5
Oxidation	5	3
Theft	5	.5

- Stationary distribution of the coverage types

Coverage type	Proportion
Breakage	.25
Breakage + Oxidation	.45
Breakage + Oxidation + Theft	.30

- Stationary distribution of the brand types

Phone brand	Proportion	Base price
Brand 1	.45	600
Brand 2	.30	550
Brand 3	.15	300
Brand 4	.10	150

- Multiplicative link between the model and its price

Model type	Multiplicative factor
0	1
1	1.15
2	1.15^2
3	1.15^3

- Claim frequencies assumptions

Coverage type	Yearly incidence
Breakage	.15
Oxidation	.05
Theft	.05 x model type

A competing model between risks is assumed

- Reporting delay hazard rate $\alpha = 0.4,\ \beta = 10$:

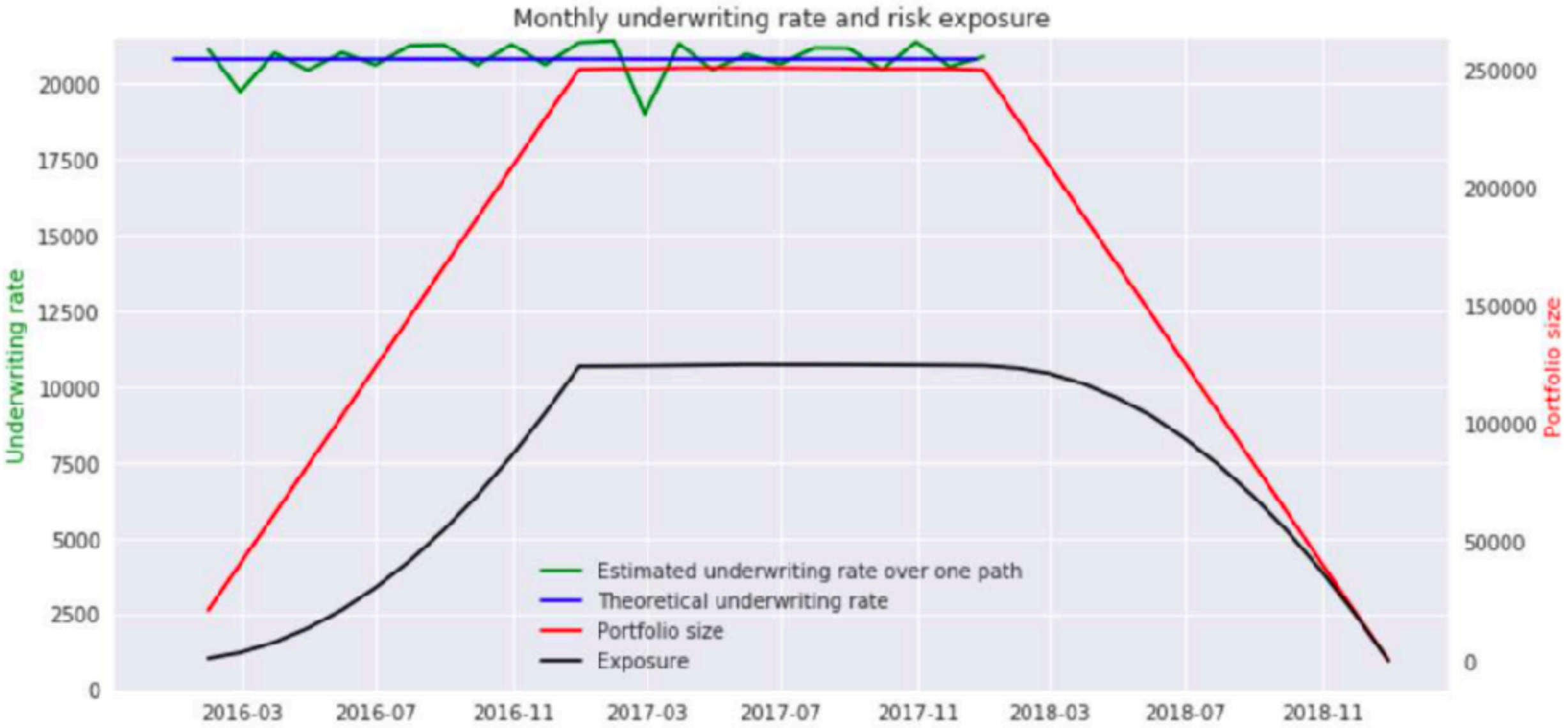
$$\lambda_{1,t+T_0} = \frac{t^{\alpha-1}\left(1-t\right)^{\beta-1}}{\int_t^1 u^{\alpha-1}\left(1-u\right)^{\beta-1} du}, \qquad 0 < t < 1.$$

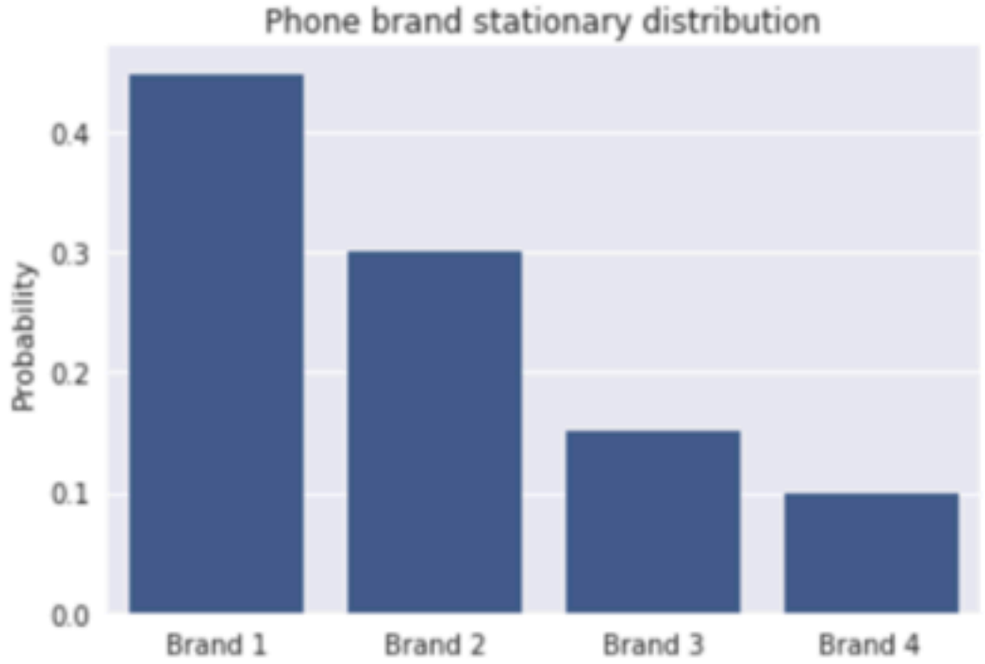
- Payment delay hazard rate $\alpha = 7,\ \beta = 7,\ m = \frac{40}{365},\ d = \frac{10}{365}$:

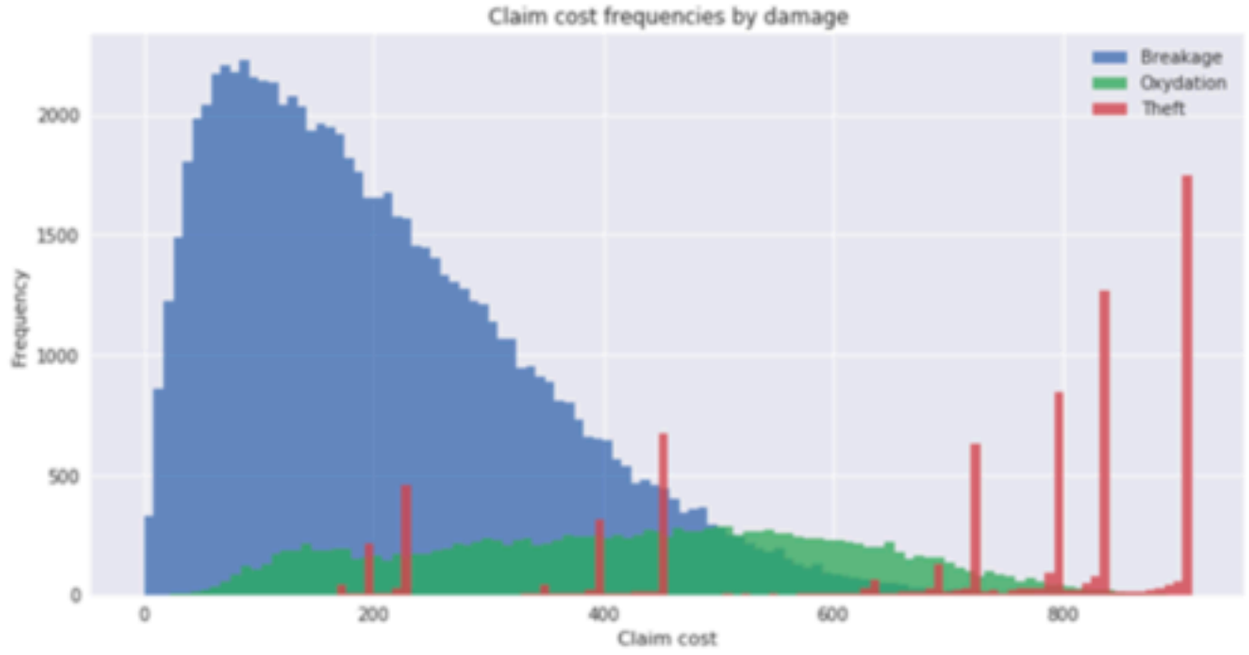
$$\lambda_{2,t+T_1} = \frac{((t-d)/m)^{\alpha-1}\left(1-(t-d)/m\right)^{\beta-1}}{m\int_{(t-d)/m}^1 u^{\alpha-1}\left(1-u\right)^{\beta-1} du}, \qquad d < t < m + d.$$

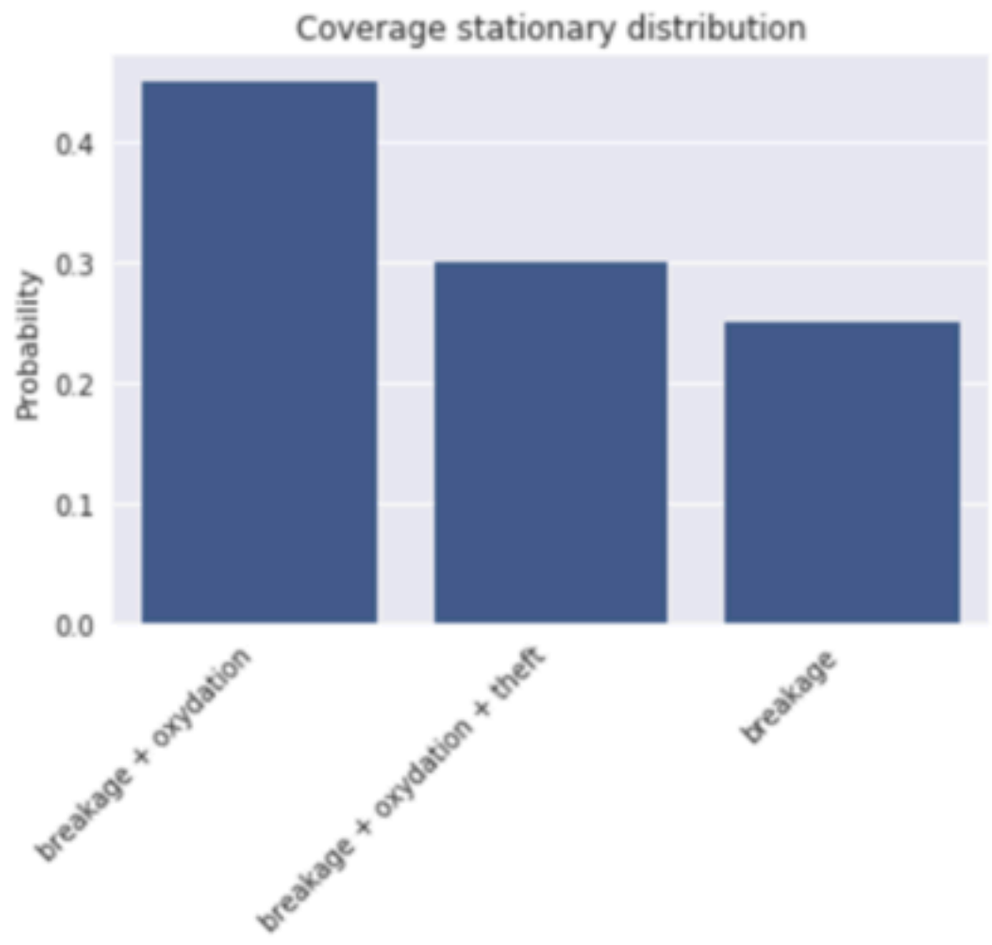
These hazard rates don't depend neither on the brand, the model, the coverage type nor the occurence date.

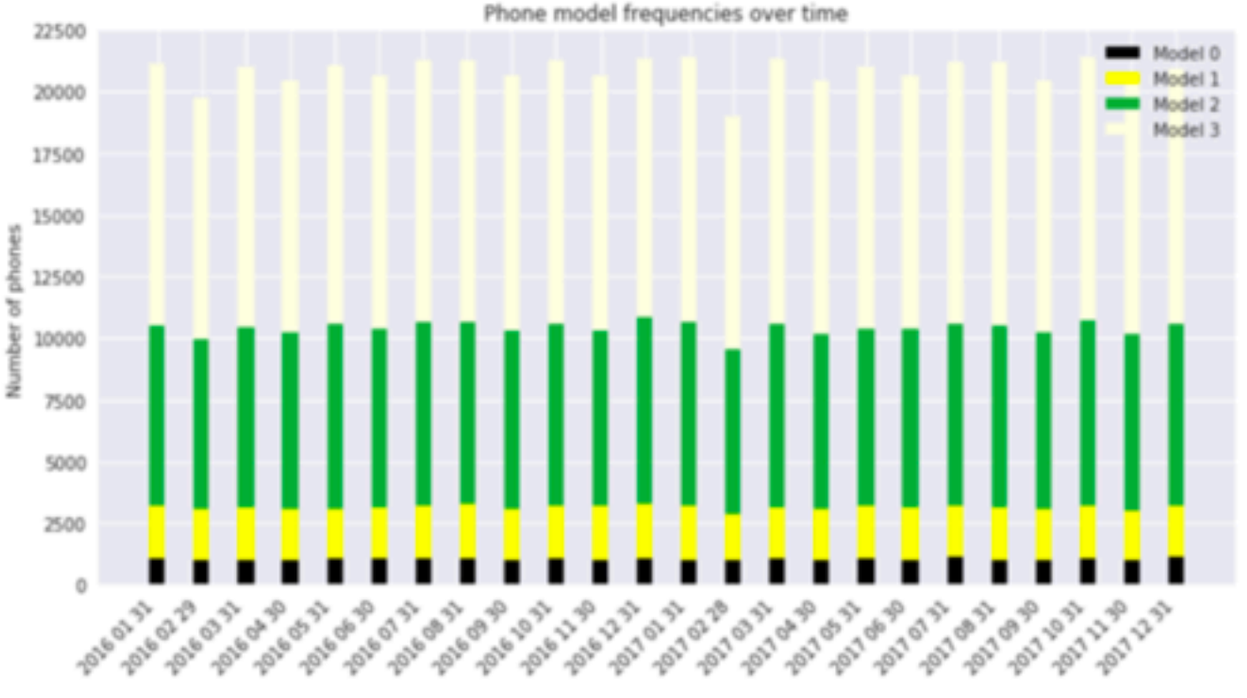
CENTRAL SCENARIO – DESCRIPTIVE STATISTICS

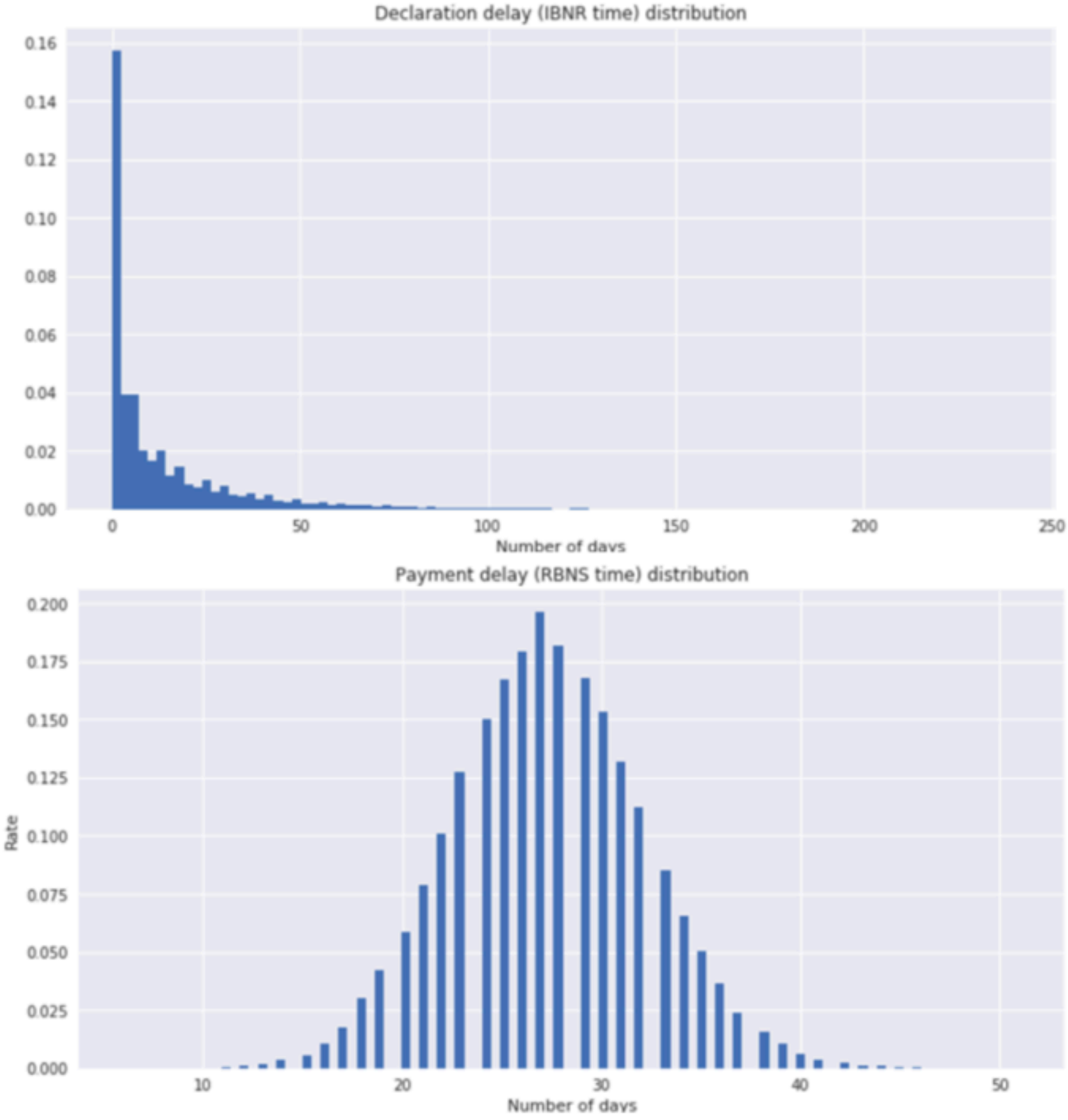



Phone brand stationary distribution
Probability
0.0
0.1
0.2
0.3
0.4
Brand 1
Brand 2
Brand 3
Brand 4


Claim cost frequencies by damage
Breakage
Oxydation
Theft
Frequency
0
500
1000
1500
2000
0
200
400
600
800
Claim cost

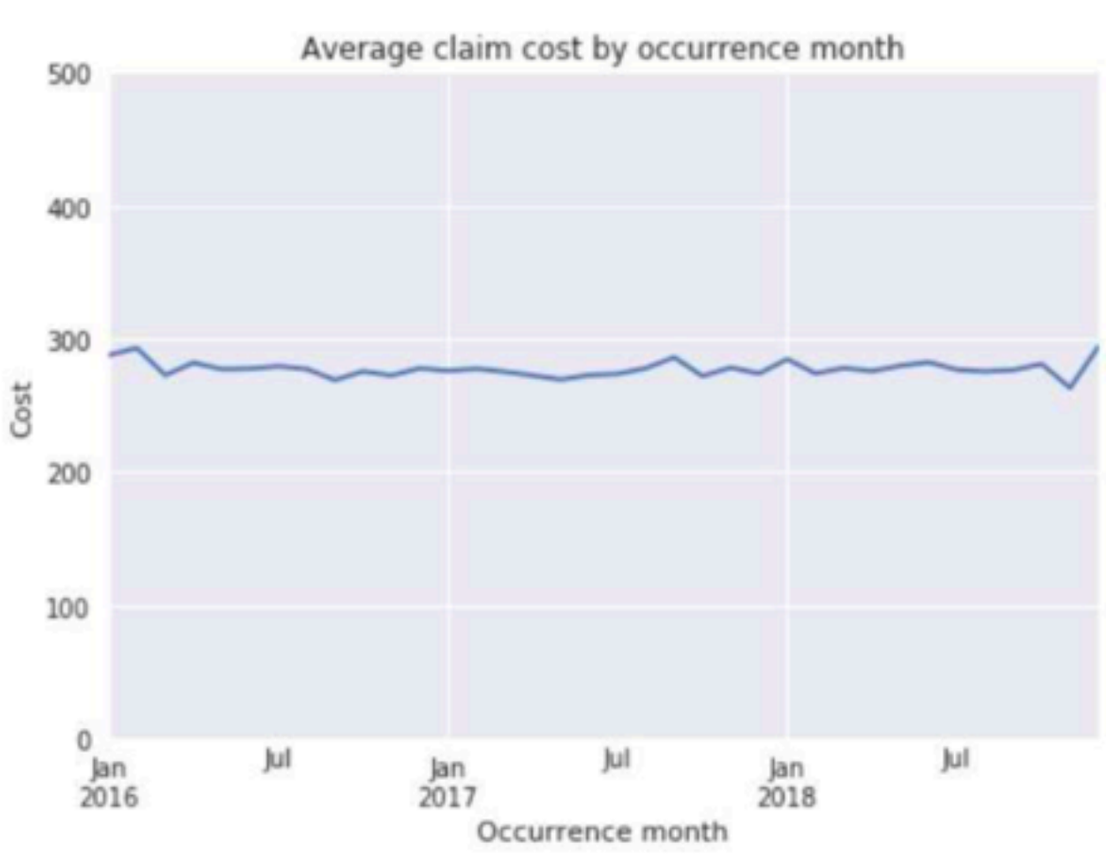

Coverage stationary distribution
Probability
0.0
0.1
0.2
0.3
0.4
breakage + oxydation
breakage + oxydation + theft
breakage

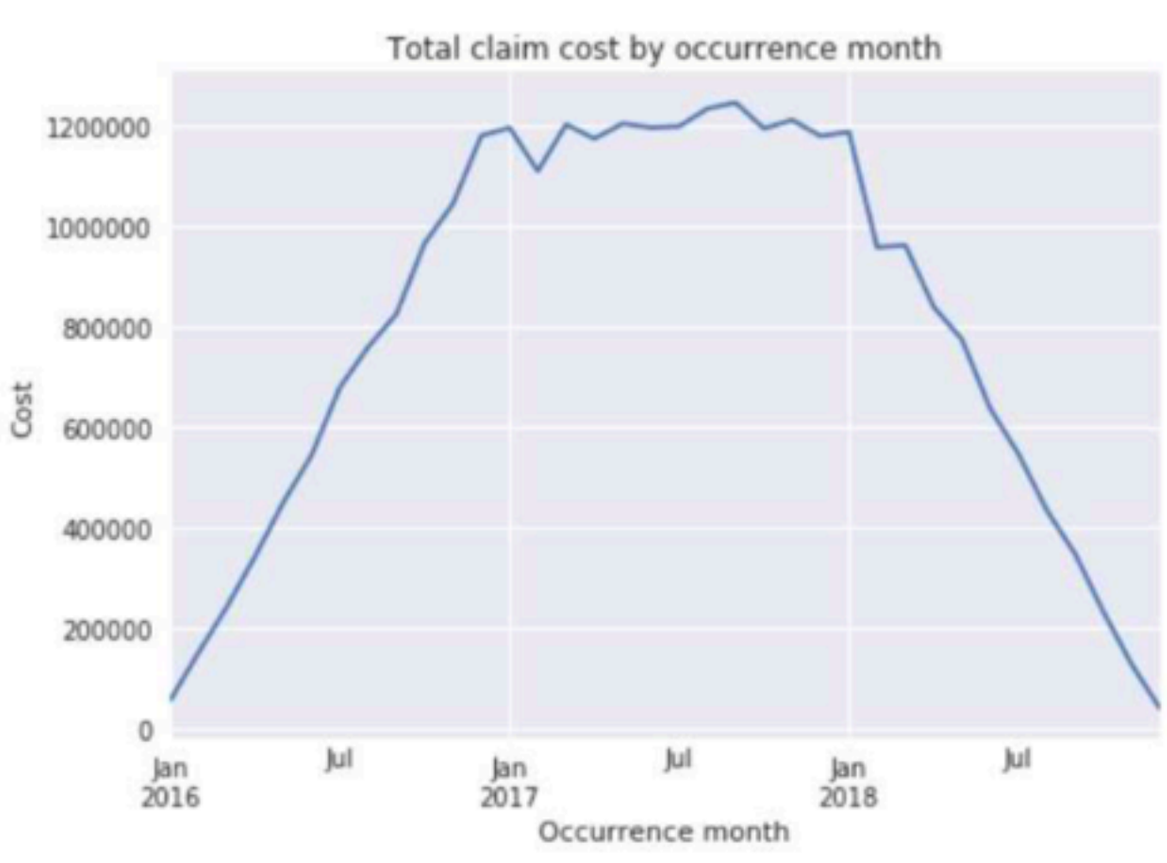

Phone model frequencies over time
Model 0
Model 1
Model 2
Model 3
Number of phones
0
2500
5000
7500
10000
12500
15000
17500
20000
22500
2016 01 31
2016 02 29
2016 03 31
2016 04 30
2016 05 31
2016 06 30
2016 07 31
2016 08 31
2016 09 30
2016 10 31
2016 11 30
2016 12 31
2017 01 31
2017 02 28
2017 03 31
2017 04 30
2017 05 31
2017 06 30
2017 07 31
2017 08 31
2017 09 30
2017 10 31
2017 11 30
2017 12 31


Declaration delay (IBNR time) distribution
0.16
0.14
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
50
100
150
200
250
Number of days
Payment delay (RBNS time) distribution
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
Rate
10
20
30
40
50
Number of days


Average monthly phone price over time by brand
Price
Underwriting date
Brand 1
Brand 2
Brand 3
Brand 4


Monthly claim rate over time by phone brand
Claim rate
Brand 1
Brand 2
Brand 3
Brand 4


Average claim cost by occurrence month
Cost
Occurrence month

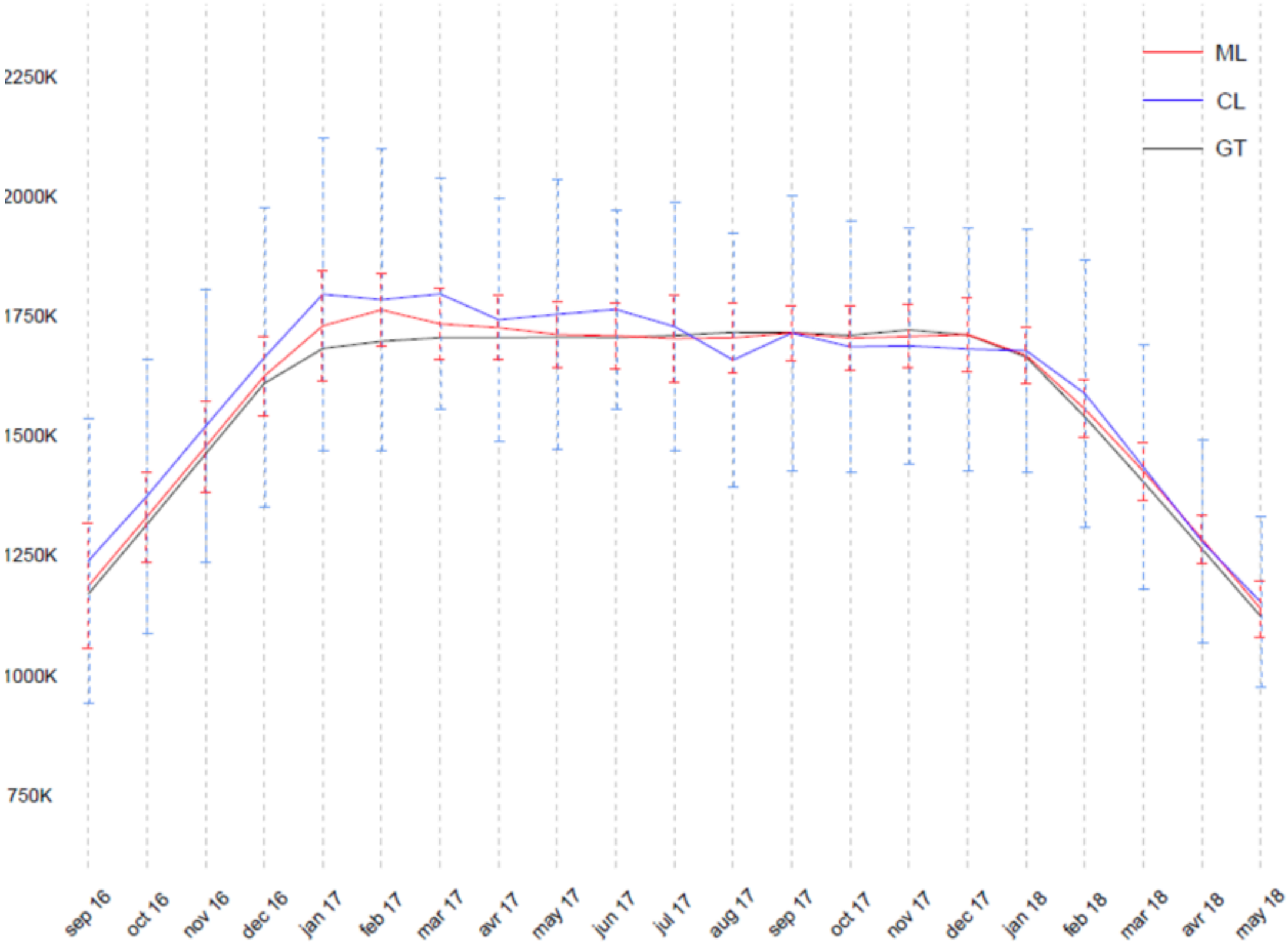

Total claim cost by occurrence month
Cost
Occurrence month

SELECTED FEATURES

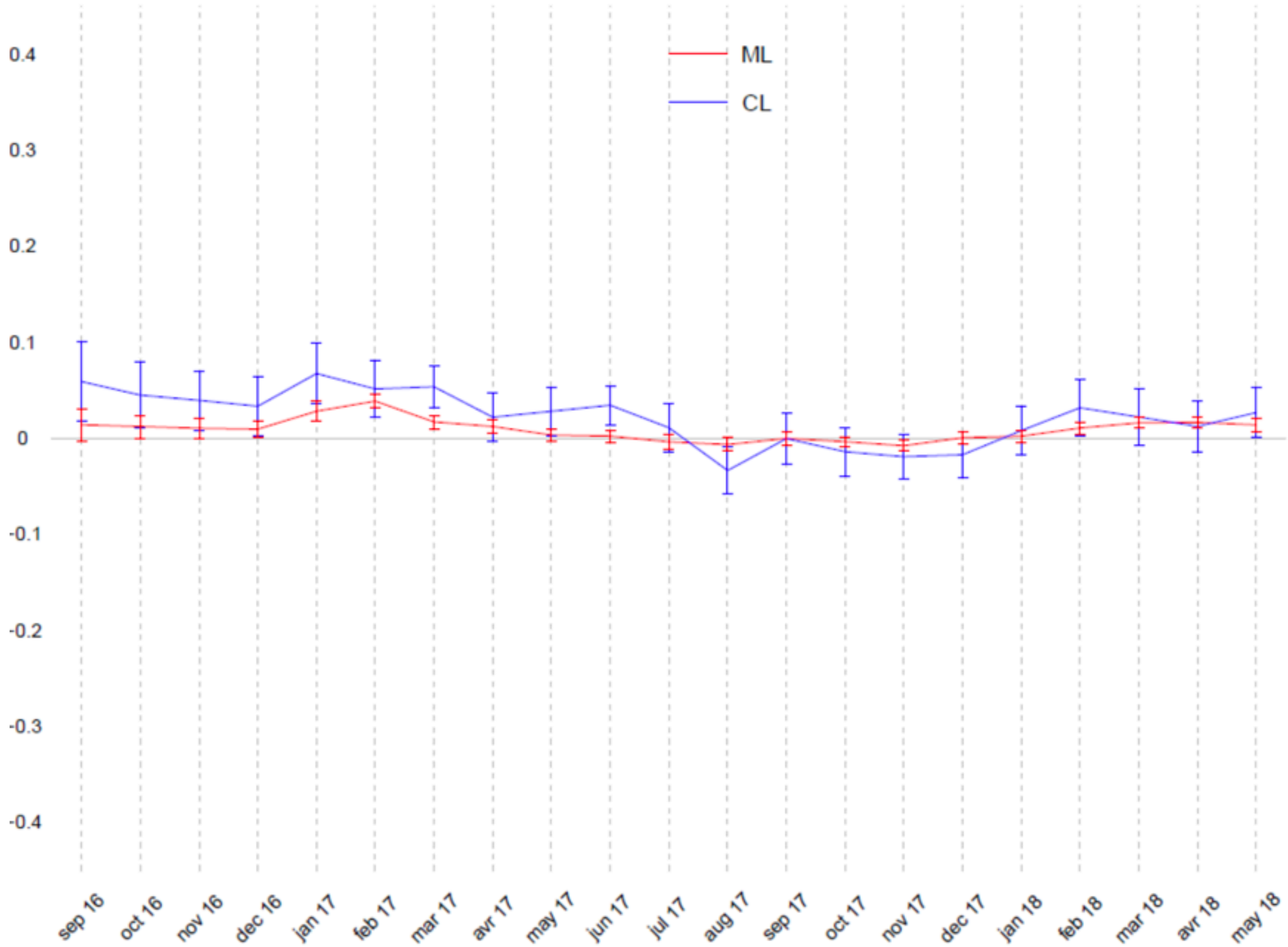
- **Features related to the contract :**
  - Brand of the mobile phone.
  - Price of the mobile phone.
  - Type of coverage (“breakage”, “breakage and oxidation” and “breakage, oxidation and theft”).
  - Underwriting date.
- **Features related to the history of the contract :**
  - Number of days since the underwriting date and exposure.
  - Indicator function whether a claim has been declared or not.
  - Type of damage (breakage, oxydation, theft).
  - Number of days since the claim has been declared.

Means of reserve predictions – central scenario

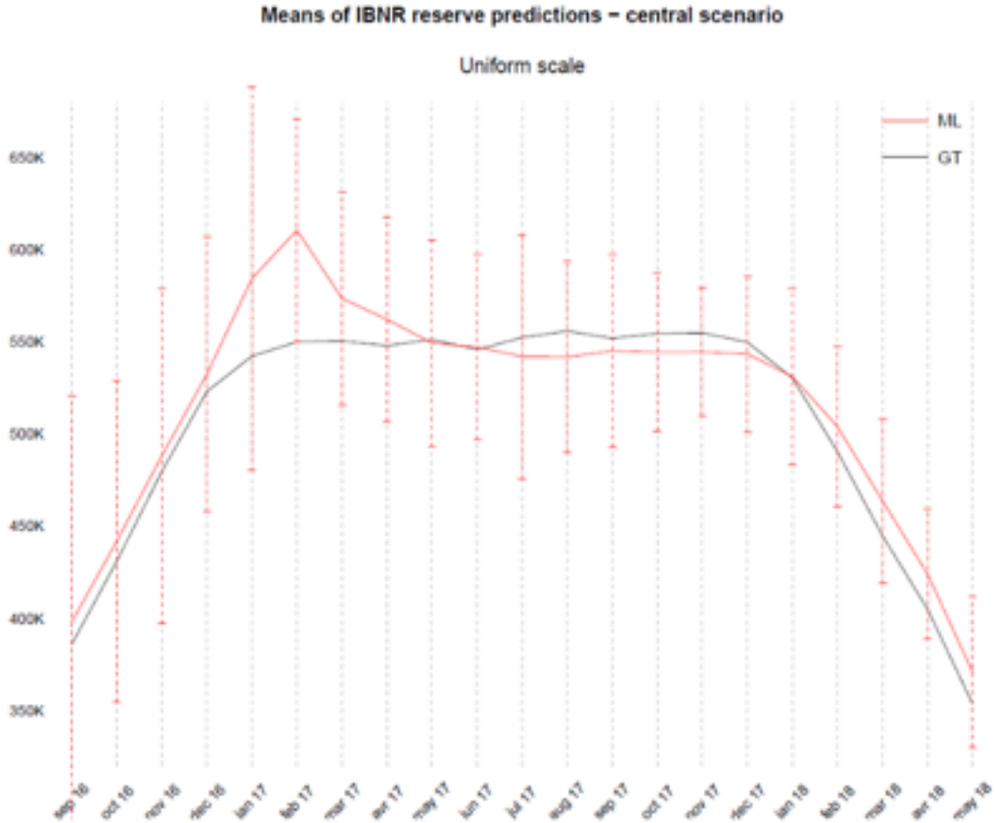
Uniform scale

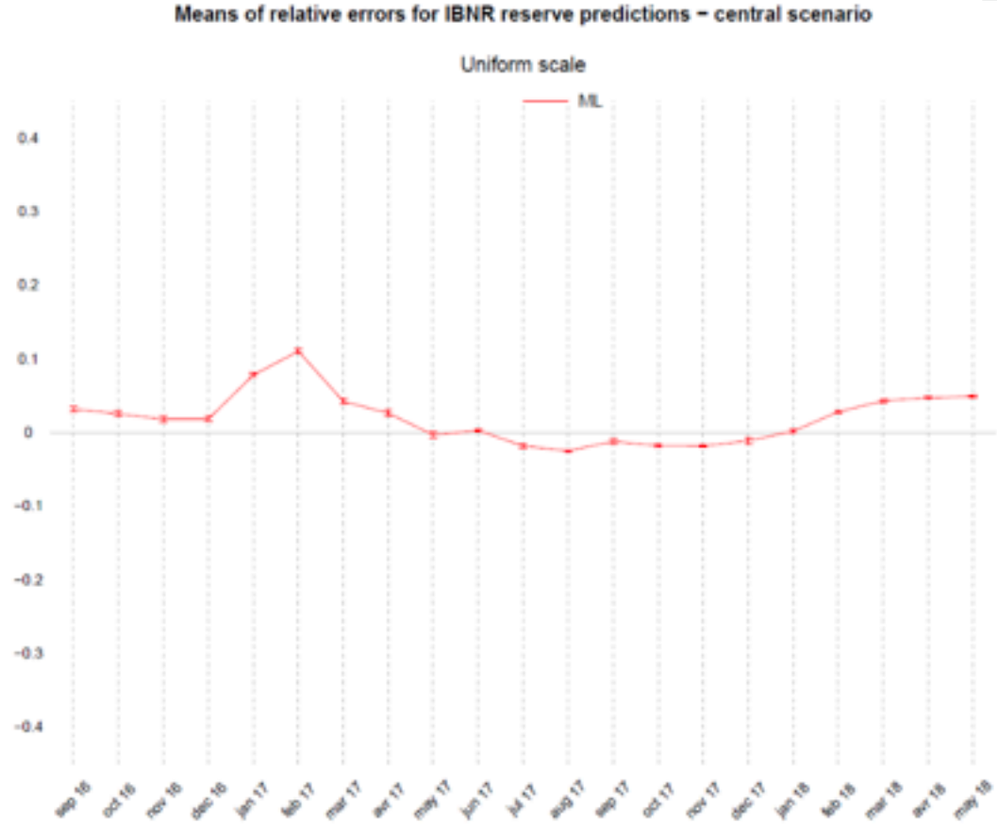


Means of relative errors for reserve predictions – central scenario

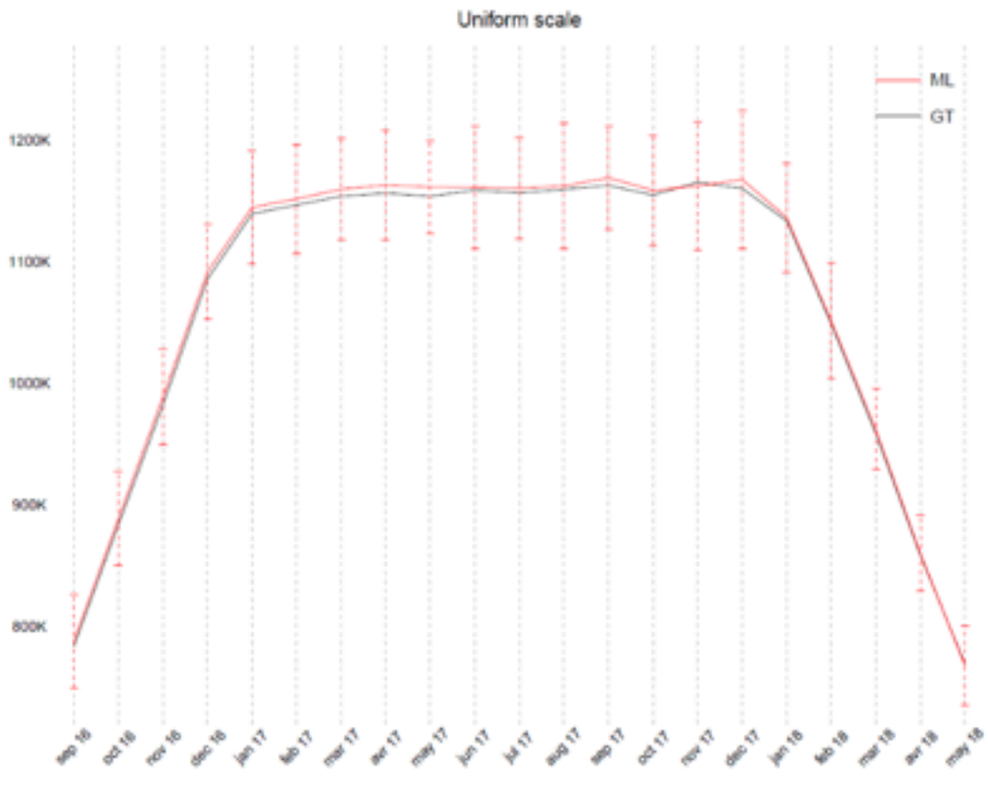


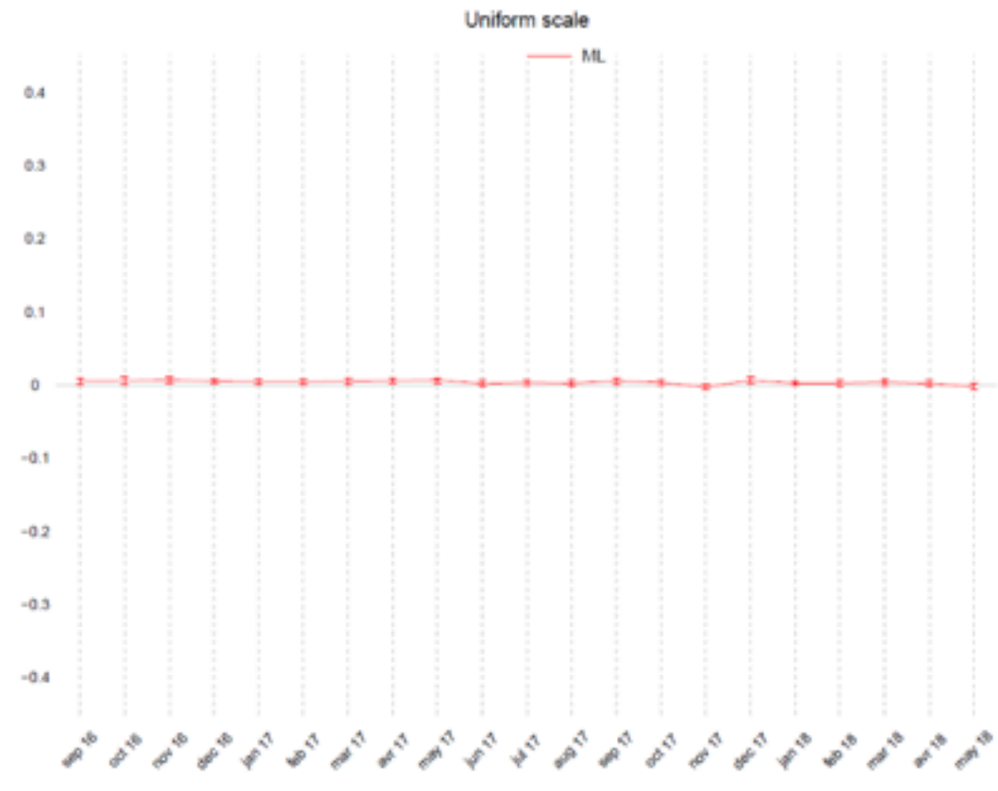
IBNR

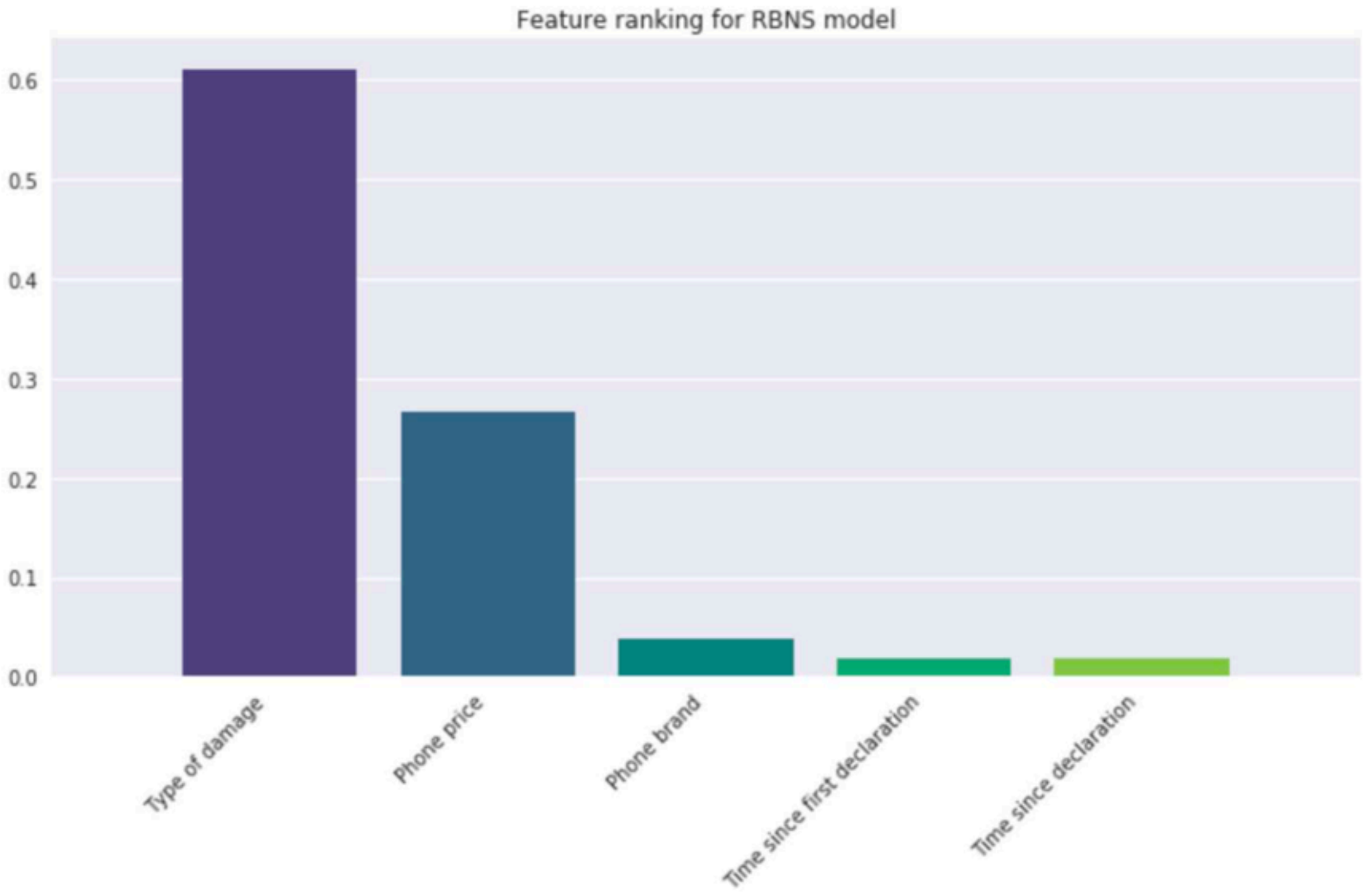

Means of IBNR reserve predictions – central scenario
Uniform scale
ML
GT
650K
600K
550K
500K
450K
400K
350K
sep 16
oct 16
nov 16
dec 16
jan 17
feb 17
mar 17
avr 17
may 17
jun 17
jul 17
aug 17
sep 17
oct 17
nov 17
dec 17
jan 18
feb 18
mar 18
avr 18
may 18

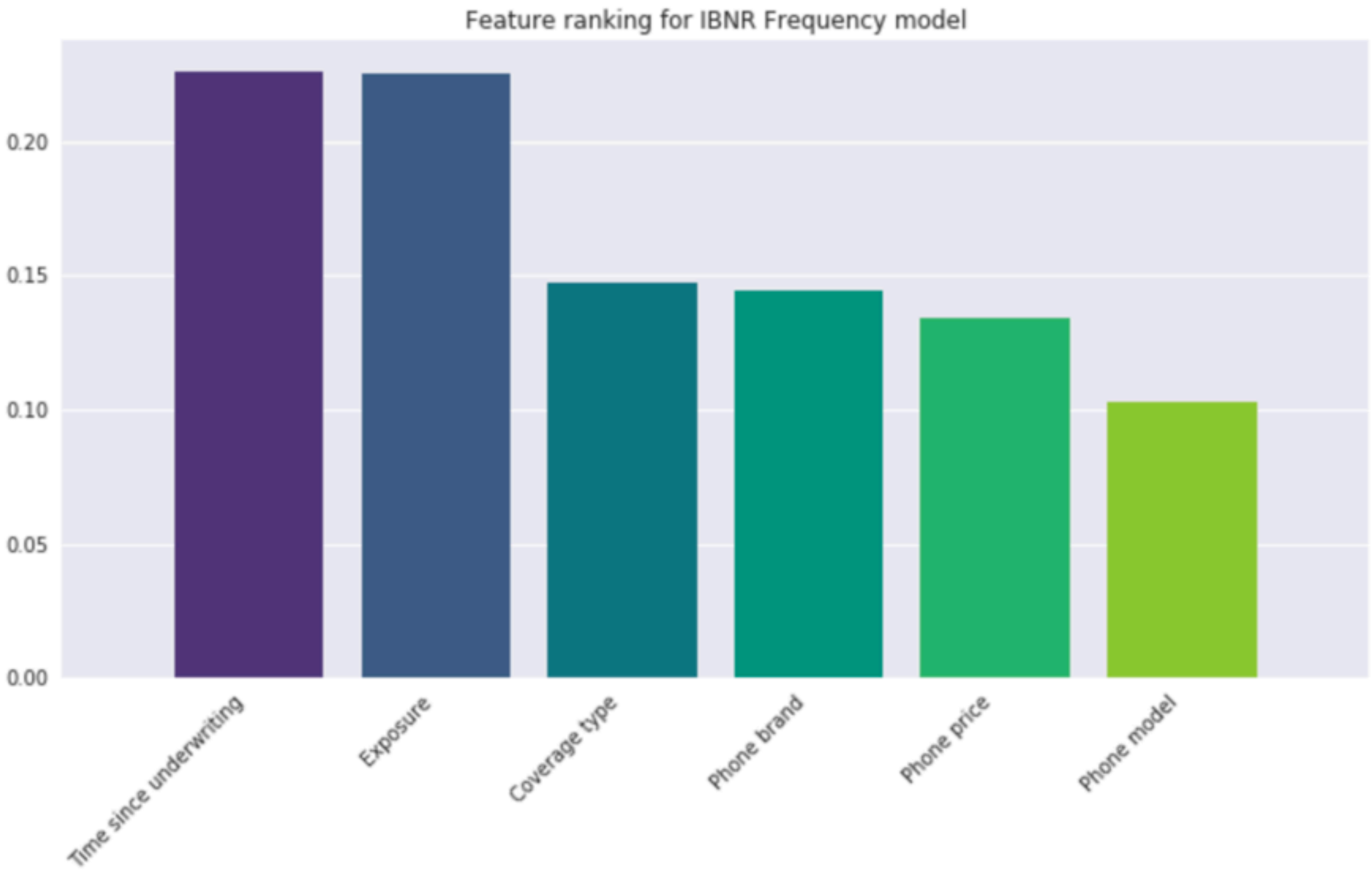

Means of relative errors for IBNR reserve predictions – central scenario
Uniform scale
ML
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
sep 16
oct 16
nov 16
dec 16
jan 17
feb 17
mar 17
avr 17
may 17
jun 17
jul 17
aug 17
sep 17
oct 17
nov 17
dec 17
jan 18
feb 18
mar 18
avr 18
may 18

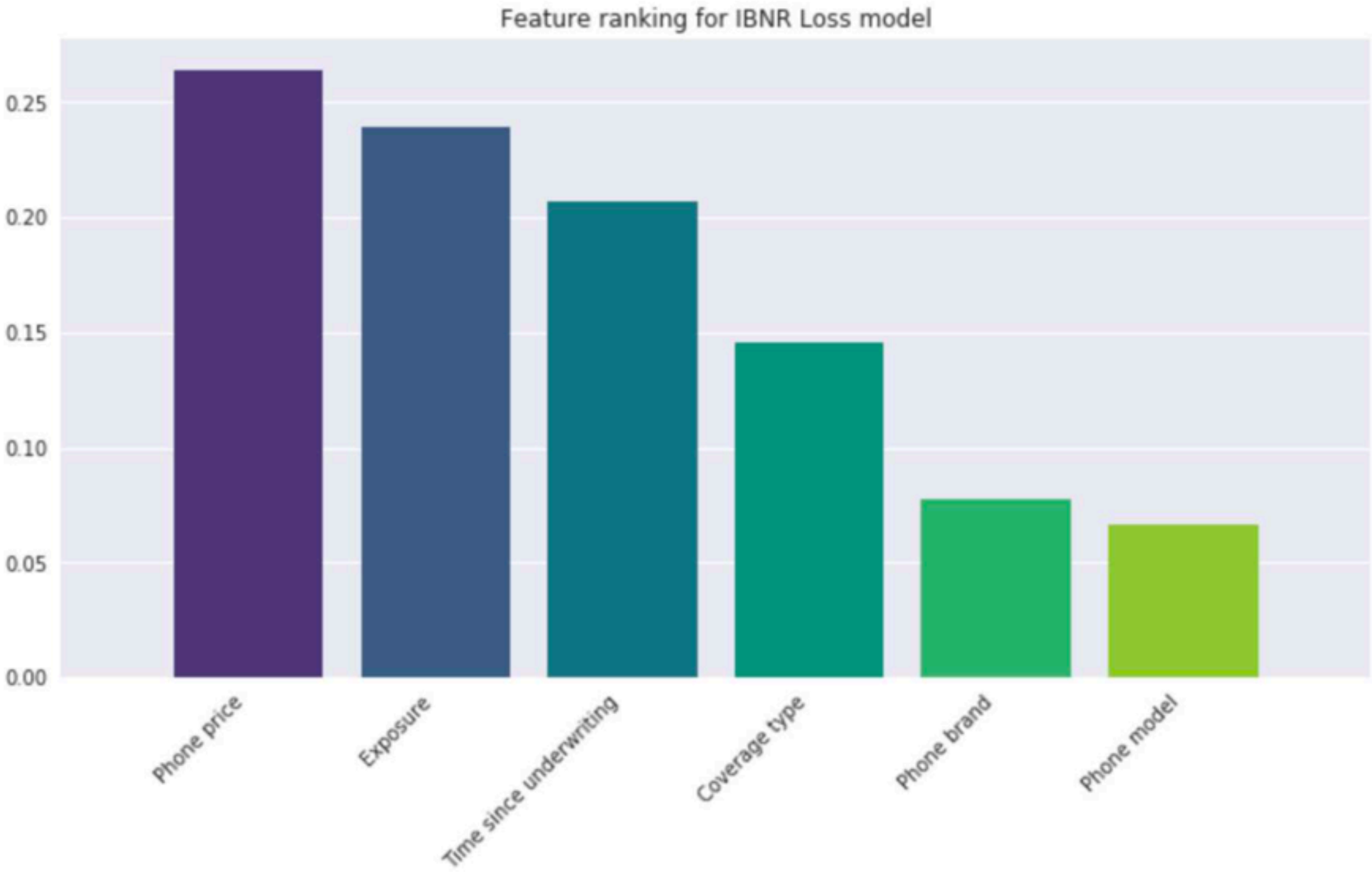
RBNS


Means of RBNS reserve predictions – central scenario
Means of RBNS reserve predictions – central scenario
Uniform scale
ML
GT
1200K
1100K
1000K
900K
800K
sep 16
oct 16
nov 16
dec 16
jan 17
feb 17
mar 17
avr 17
may 17
jun 17
jul 17
aug 17
sep 17
oct 17
nov 17
dec 17
jan 18
feb 18
mar 18
avr 18
may 18


Means of relative errors for RBNS reserve predictions – central scenario
Means of relative errors for RBNS reserve predictions – central scenario
Uniform scale
ML
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
sep 16
oct 16
nov 16
dec 16
jan 17
feb 17
mar 17
avr 17
may 17
jun 17
jul 17
aug 17
sep 17
oct 17
nov 17
dec 17
jan 18
feb 18
mar 18
avr 18
may 18


Feature ranking for RBNS model
0.6
0.5
0.4
0.3
0.2
0.1
0.0
Type of damage
Phone price
Phone brand
Time since first declaration
Time since declaration


Feature ranking for IBNR Frequency model
0.20
0.15
0.10
0.05
0.00
Time since underwriting
Exposure
Coverage type
Phone brand
Phone price
Phone model

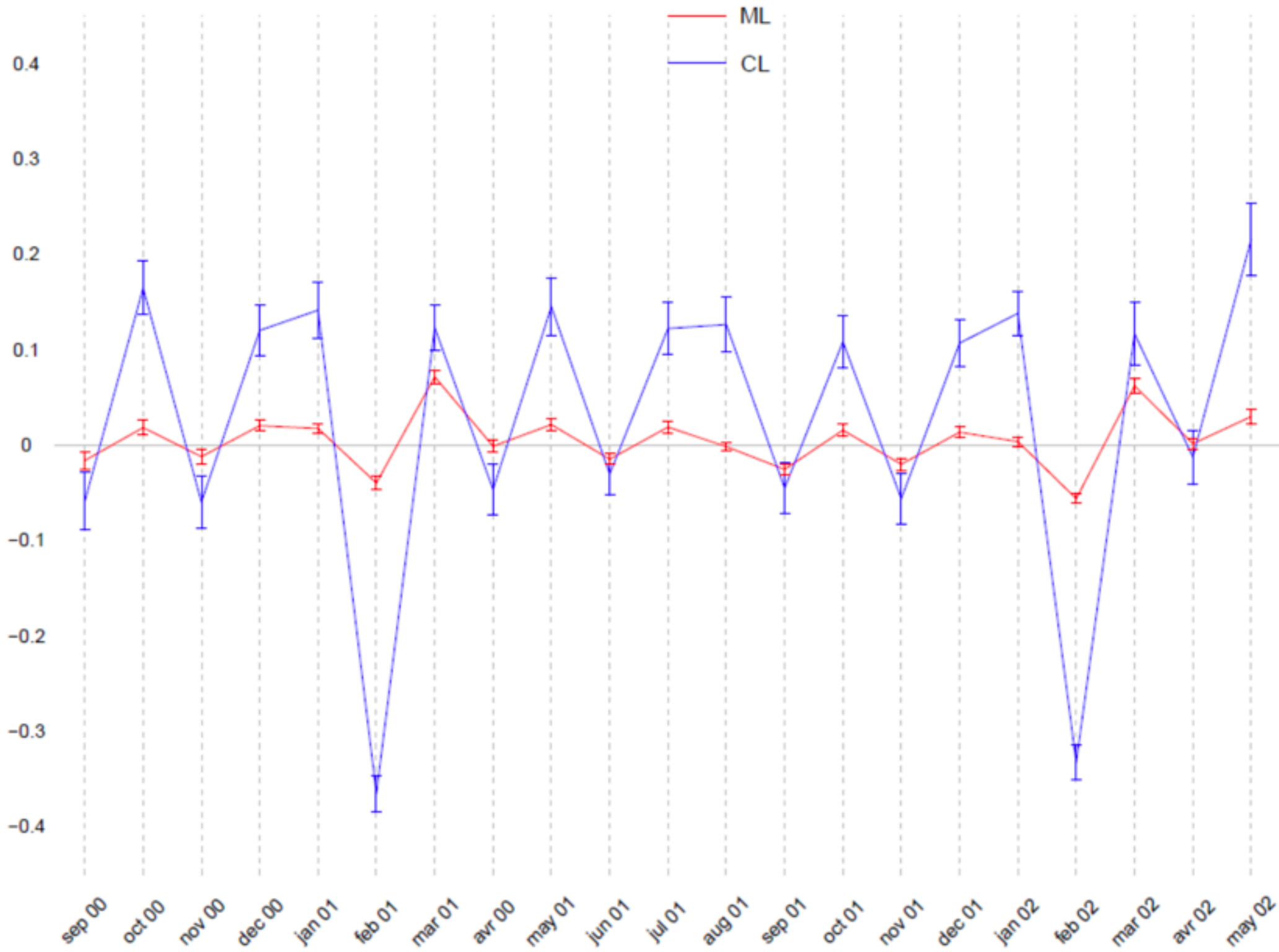

Feature ranking for IBNR Loss model
0.25
0.20
0.15
0.10
0.05
0.00
Phone price
Exposure
Time since underwriting
Coverage type
Phone brand
Phone model

OTHER SCENARII

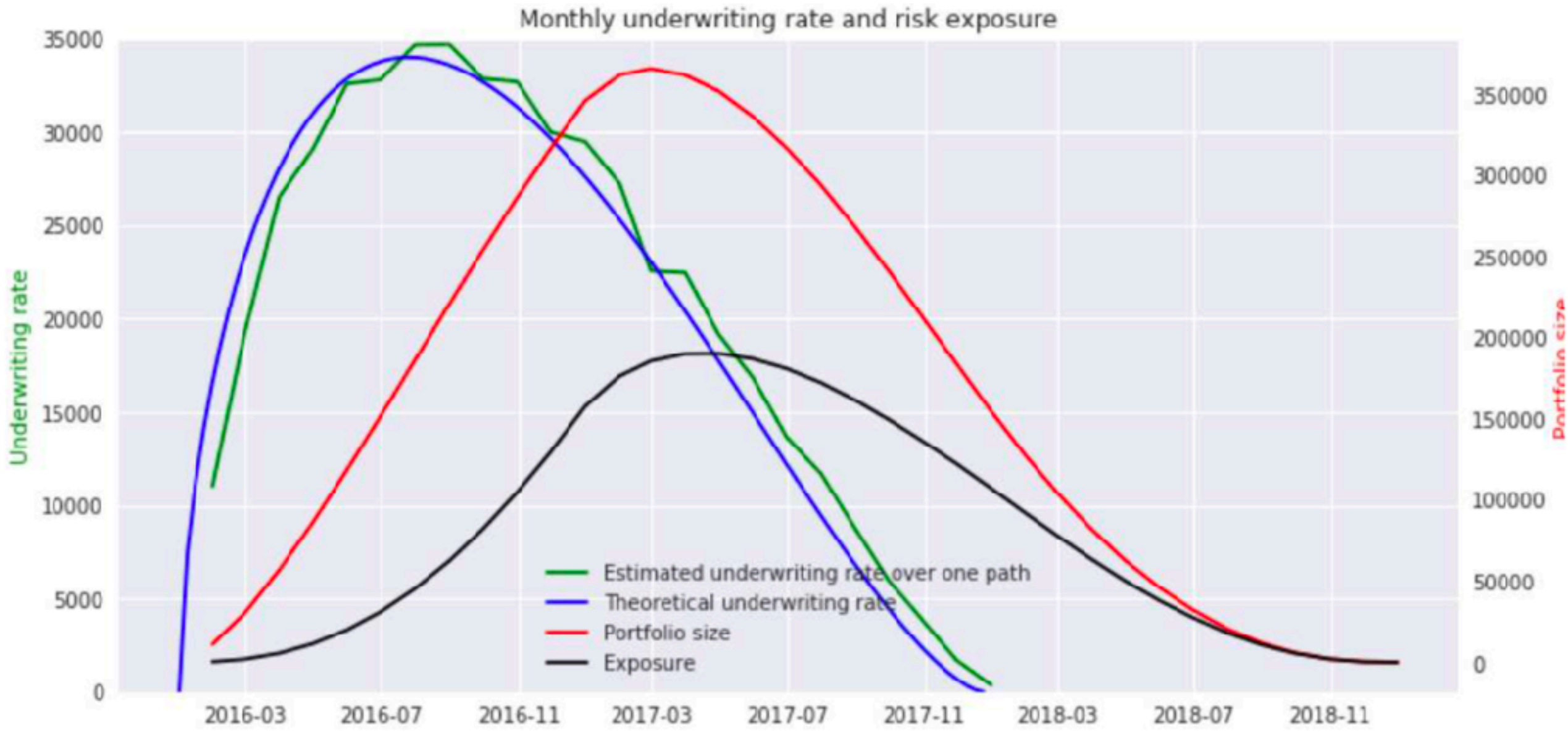
- We start from **central scenario**, then **add one of the following**:
  - Monthly scale instead of uniform scale.
  - Time-dependent underwriting rate.
  - Decrease of 10% of the payment delay since January 1st 2017.
  - Increase of 10% of the payment delay since January 1st 2017.
  - Arrivals of new and more expensive mobile phones at the end of the year 2016.
  - Increase of 40% of the claim rate from December 15th 2016 to January 15th 2017.

MONTHLY SCALE SCENARIO

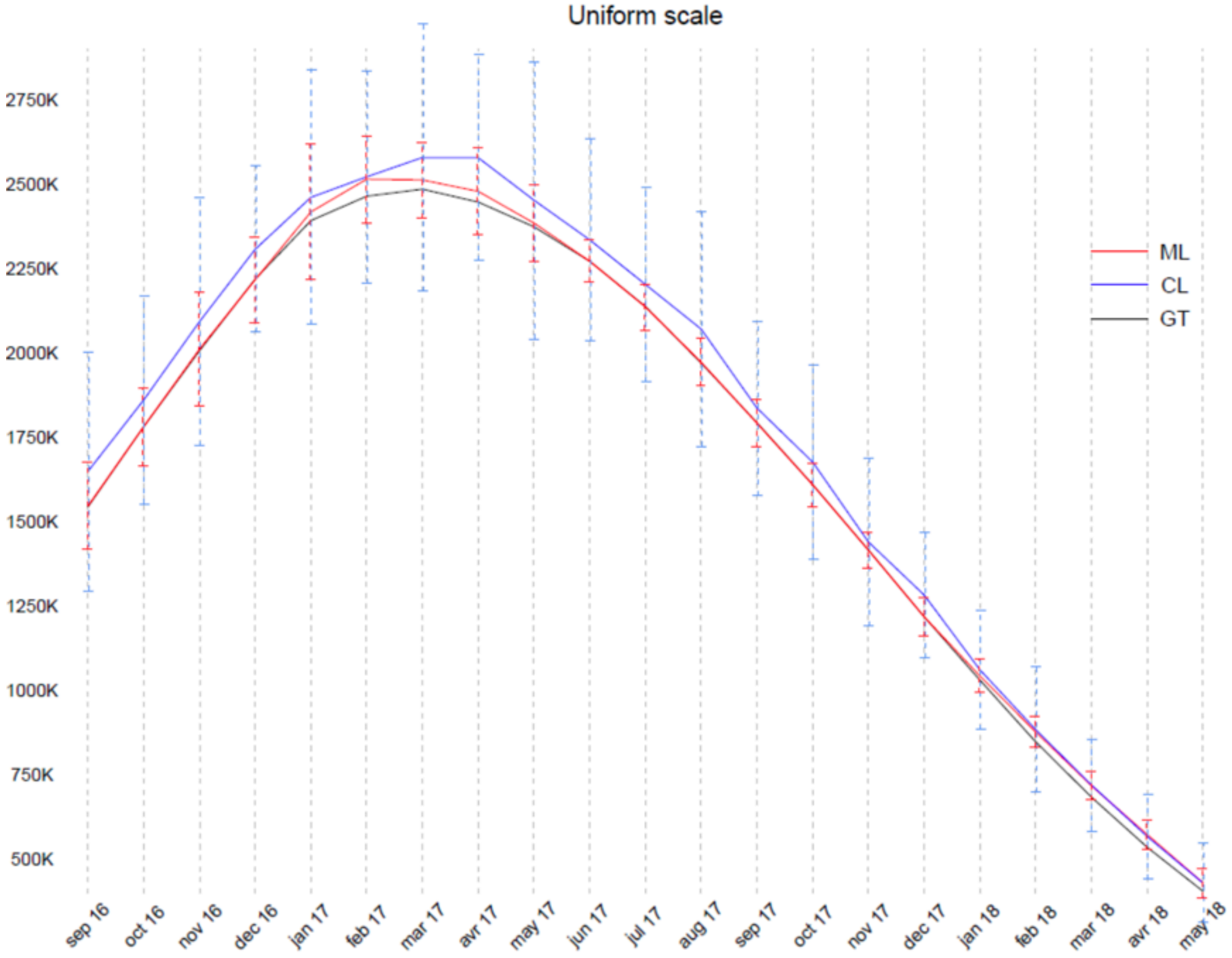
Mean errors of reserve predictions – central scenario



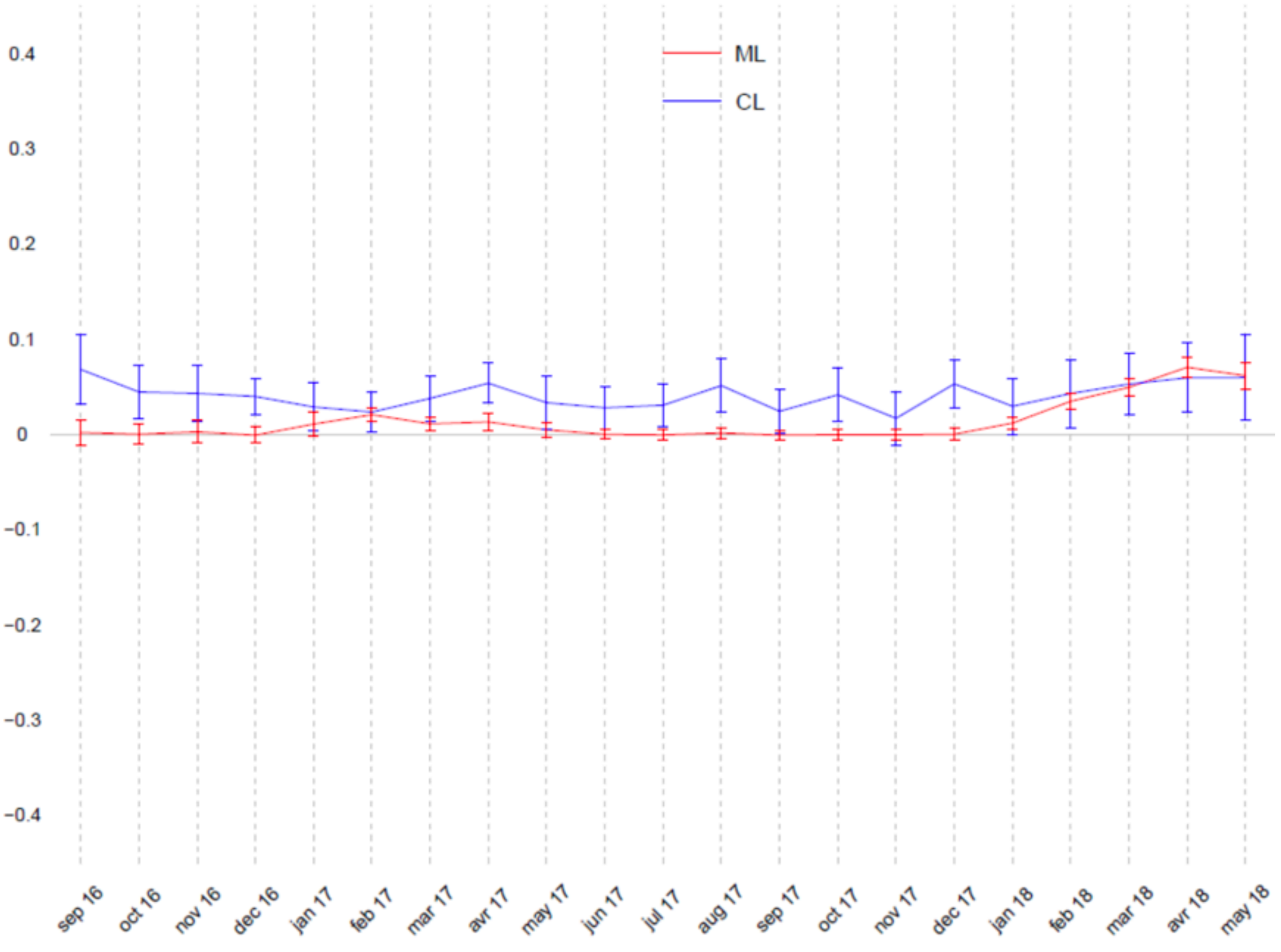
NON-CONSTANT UNDERWRITING RATE SCENARIO


Monthly underwriting rate and risk exposure
Underwriting rate
Portfolio size
35000
30000
25000
20000
15000
10000
5000
0
350000
300000
250000
200000
150000
100000
50000
0
2016-03
2016-07
2016-11
2017-03
2017-07
2017-11
2018-03
2018-07
2018-11
Estimated underwriting rate over one path
Theoretical underwriting rate
Portfolio size
Exposure

Means of reserve predictions – non-constant underwriting rate



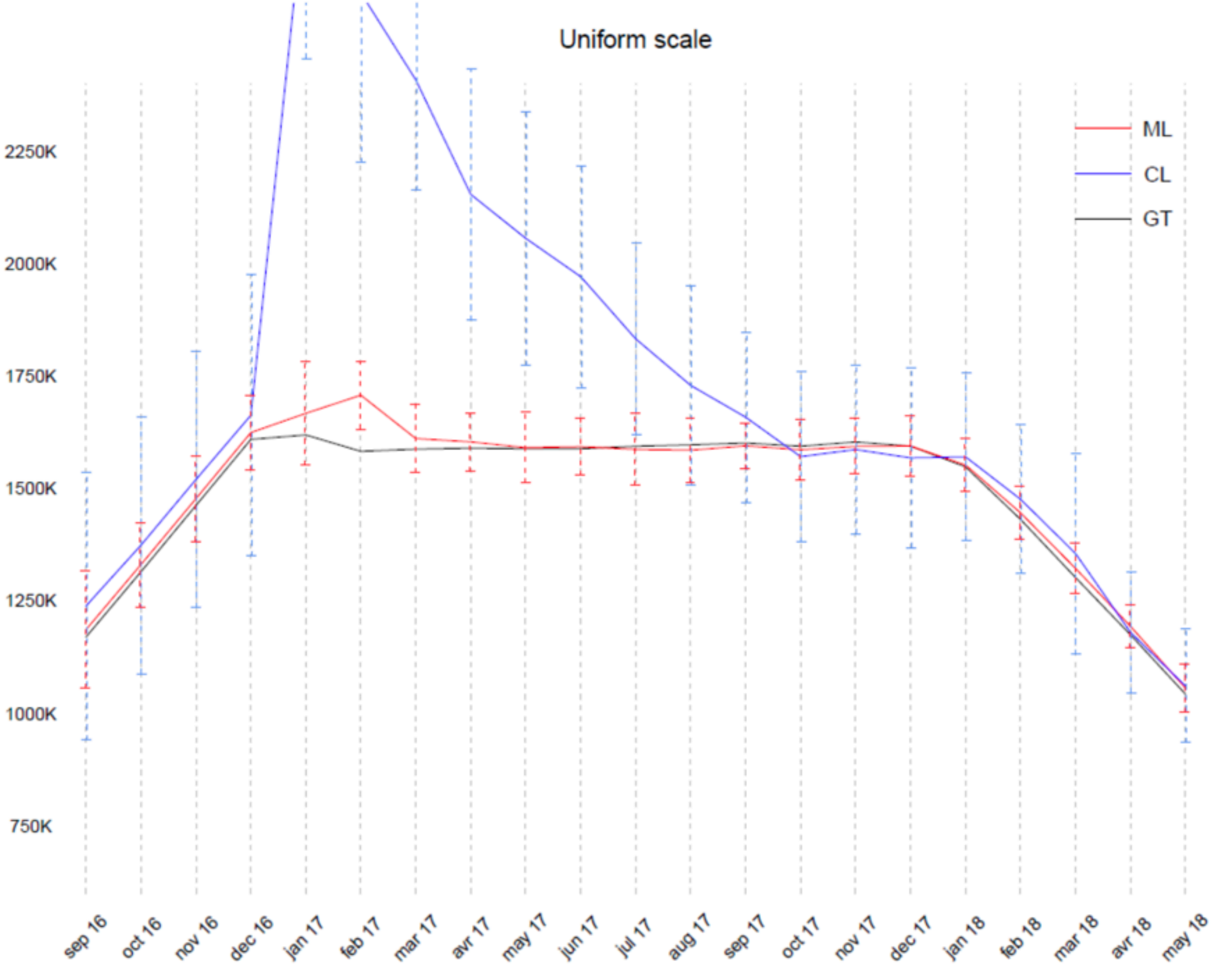
Means of relative errors for reserve predictions – non–constant underwriting rate



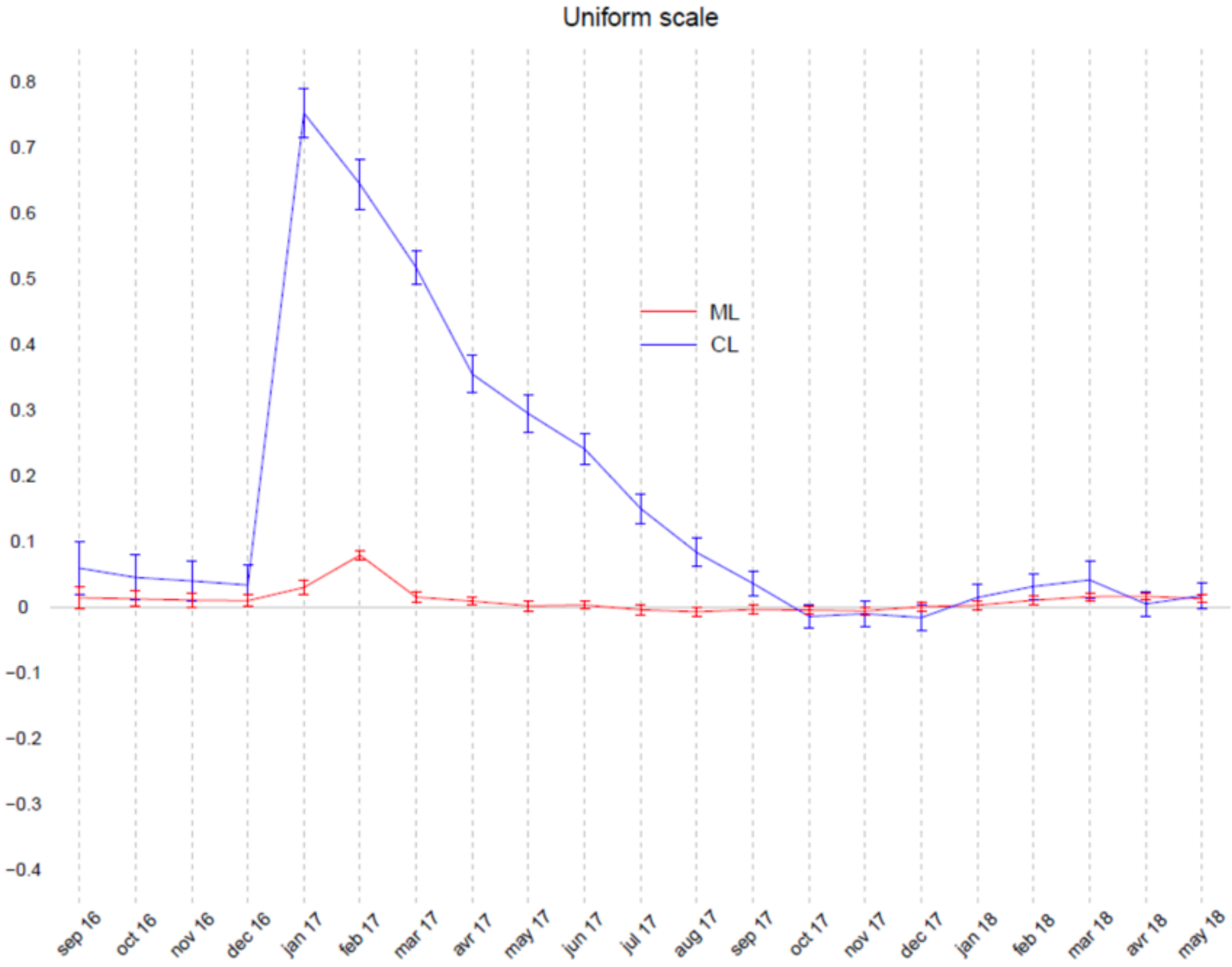
SHOCK ON PAYMENT DELAY: 10% DECREASE


Average payment delay by payment months
Average payment delay
45
40
35
30
25
20
15
2016-04
2016-08
2016-12
2017-04
2017-08
2017-12
2018-04
2018-08
2018-12
2019-04

Means of reserve predictions – shock on the payment delay



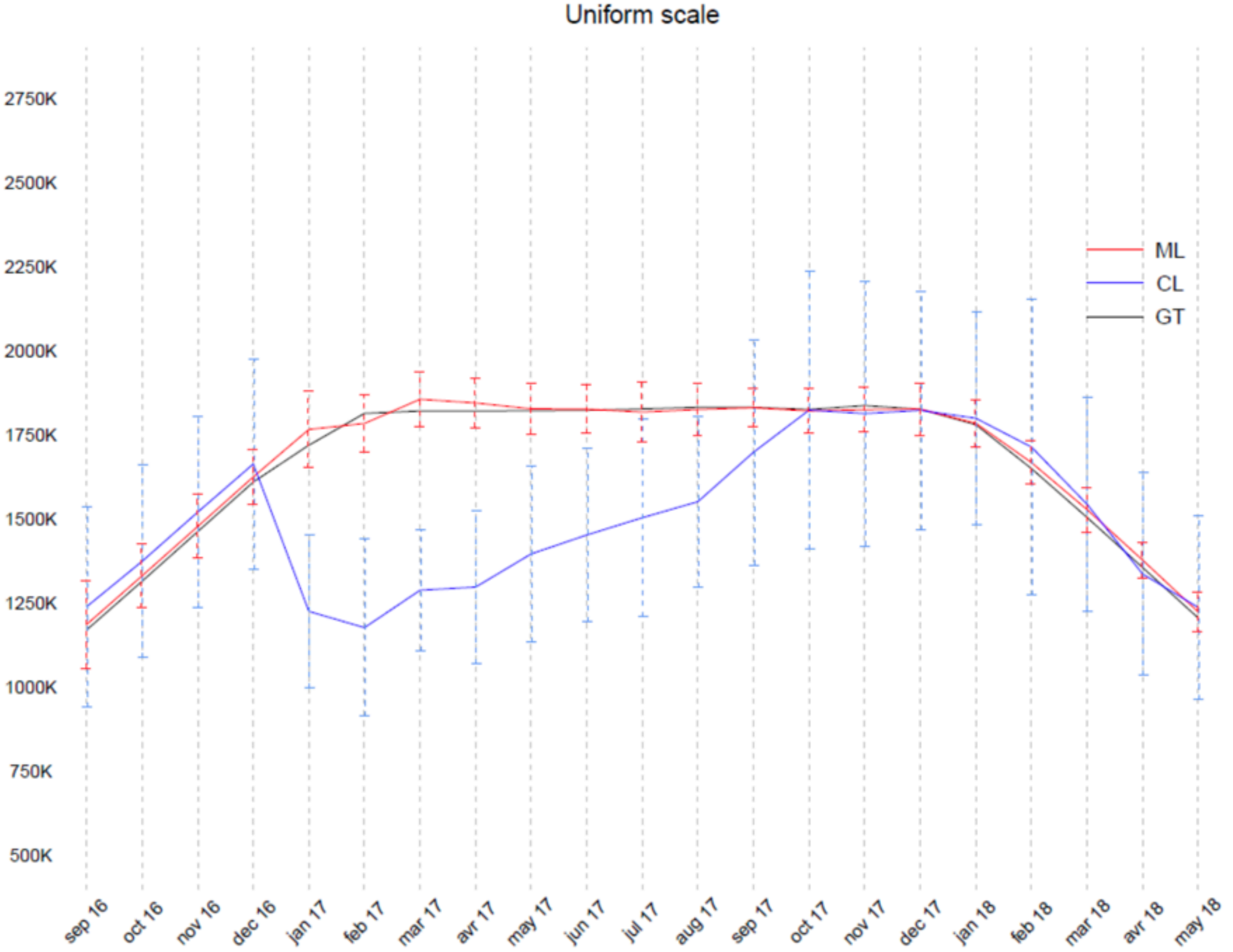
Means of relative errors for reserve predictions – shock on the payment delay



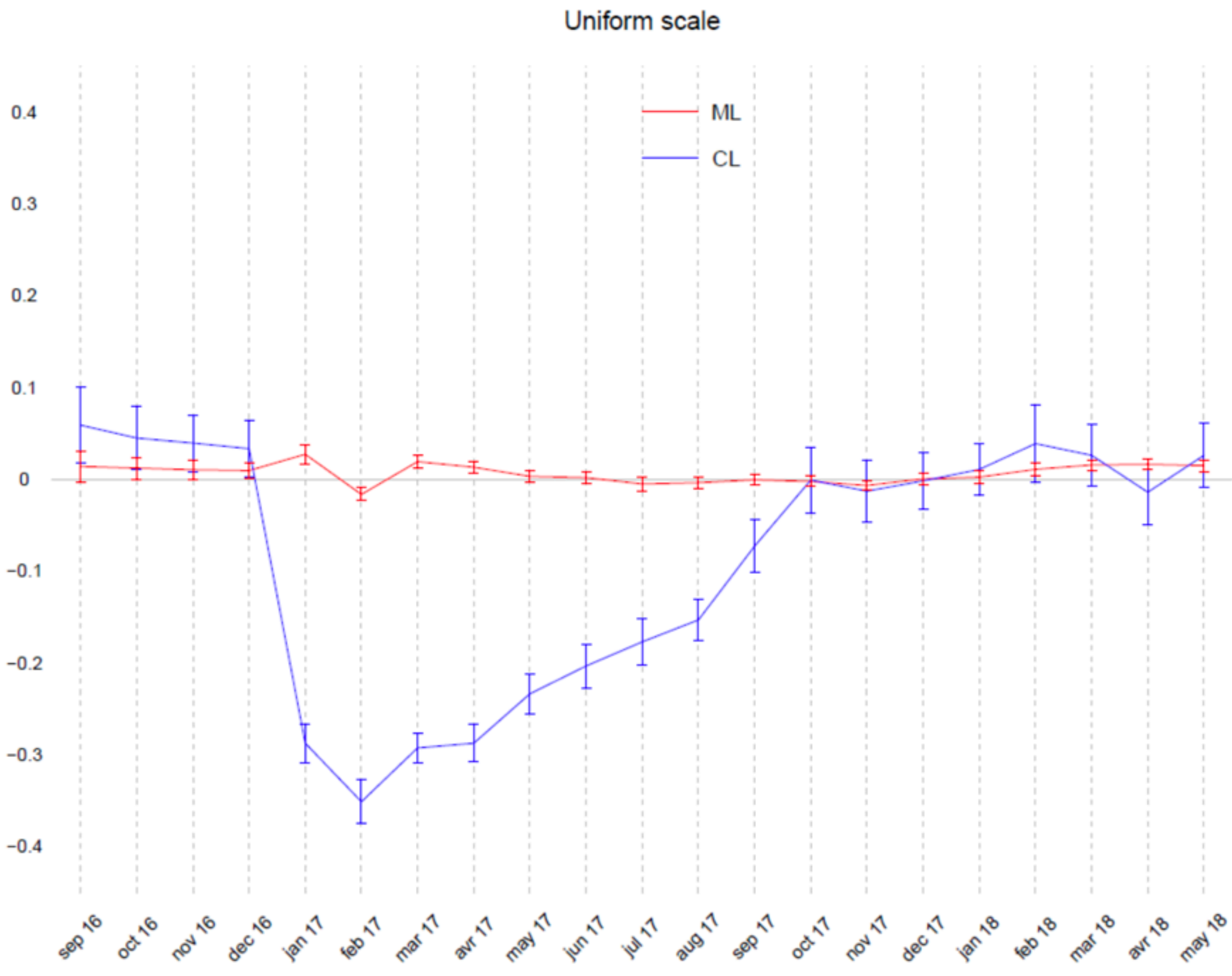
SHOCK ON PAYMENT DELAY: 10% INCREASE


Average payment delay by payment months
Average payment delay
45
40
35
30
25
20
15
2016-04
2016-08
2016-12
2017-04
2017-08
2017-12
2018-04
2018-08
2018-12
2019-04

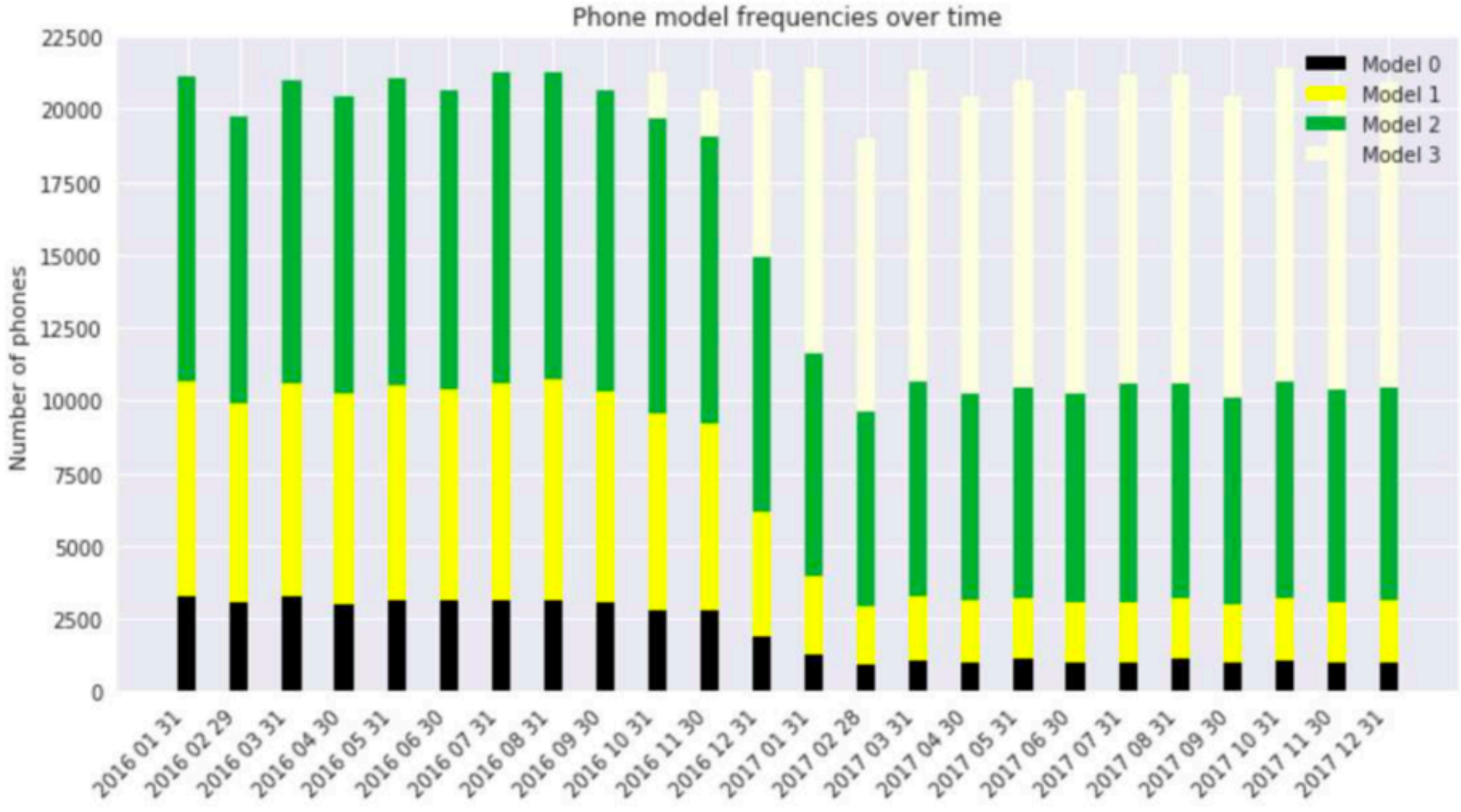
Means of reserve predictions – positive shock on the payment delay



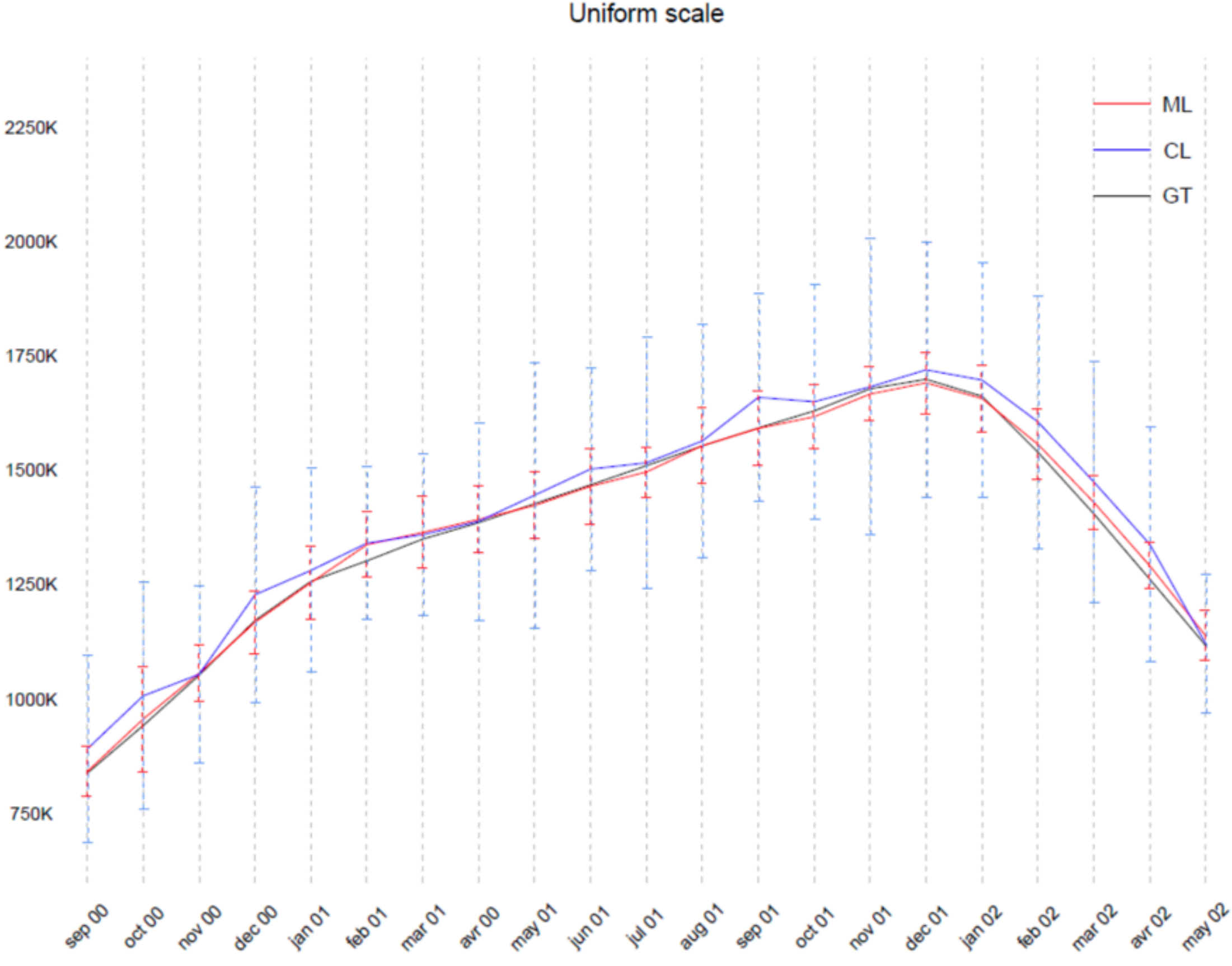
Means of relative errors for reserve predictions − positive shock on the payment delay



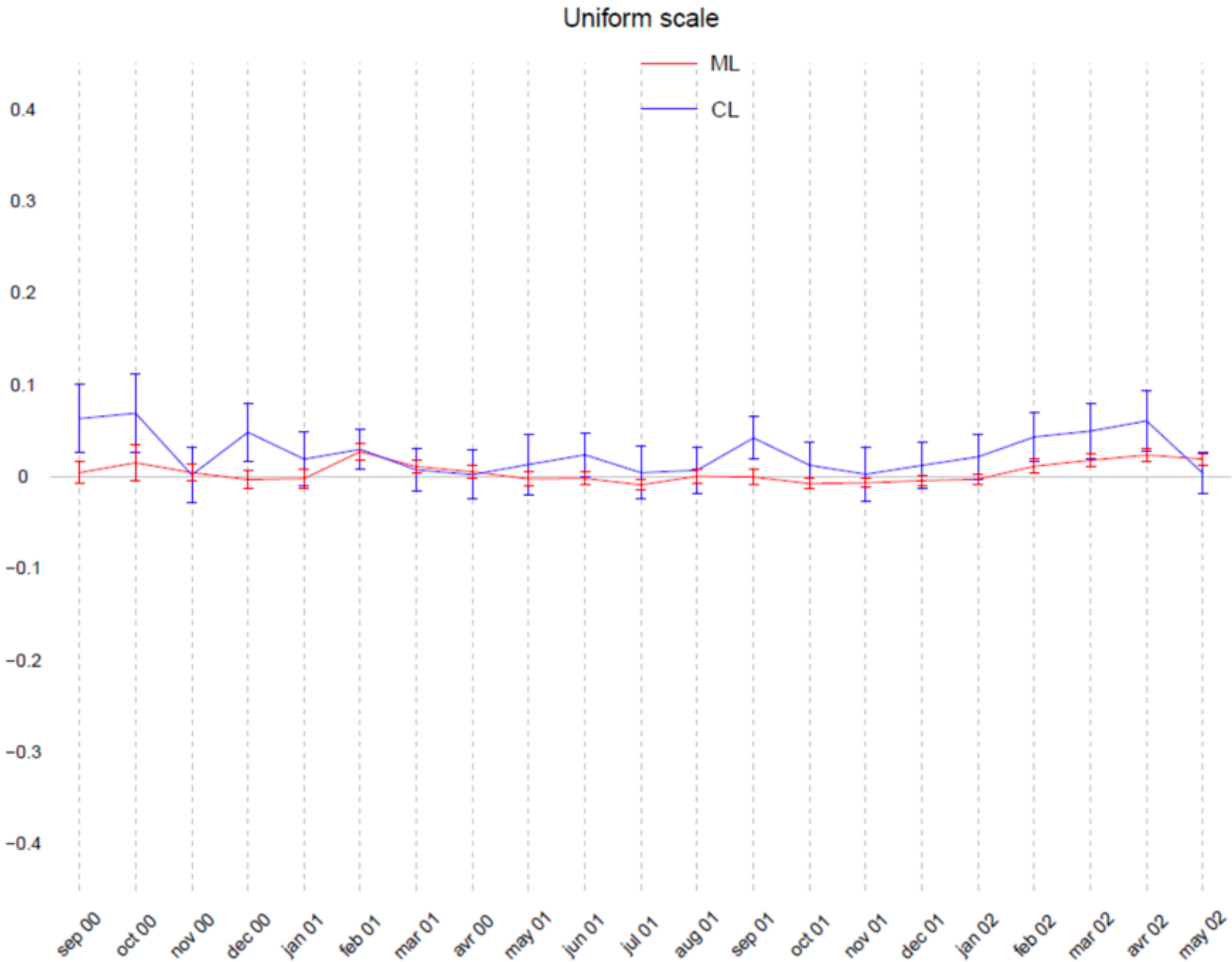
ARRIVAL OF NEW PHONE MODELS


Phone model frequencies over time
Number of phones
22500
20000
17500
15000
12500
10000
7500
5000
2500
0
Model 0
Model 1
Model 2
Model 3
2016 01 31
2016 02 29
2016 03 31
2016 04 30
2016 05 31
2016 06 30
2016 07 31
2016 08 31
2016 09 30
2016 10 31
2016 11 30
2016 12 31
2017 01 31
2017 02 28
2017 03 31
2017 04 30
2017 05 31
2017 06 30
2017 07 31
2017 08 31
2017 09 30
2017 10 31
2017 11 30
2017 12 31

Means of reserve predictions – arrivals of new mobile phones

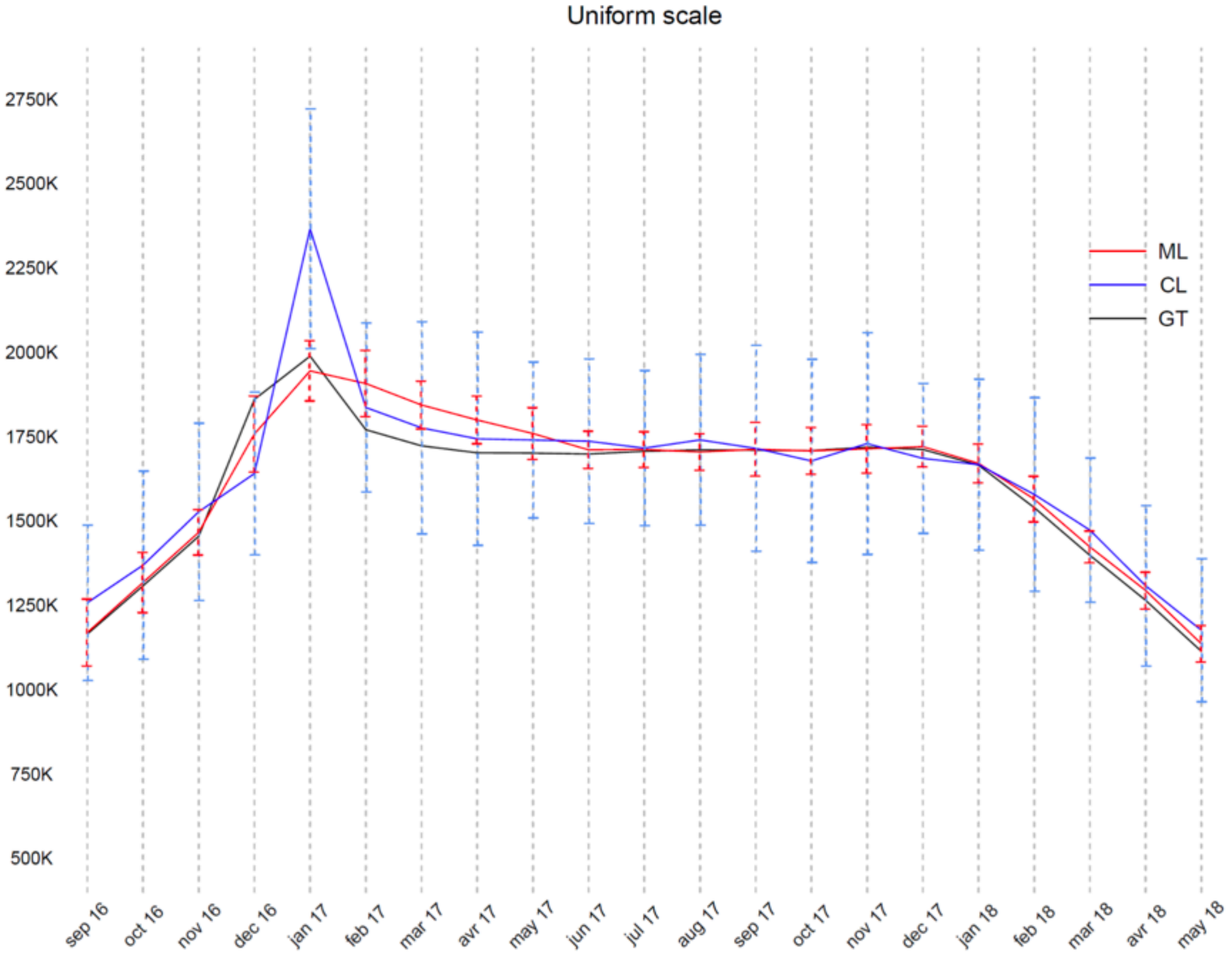


Means of relative errors for reserve predictions – arrivals of new mobile phones

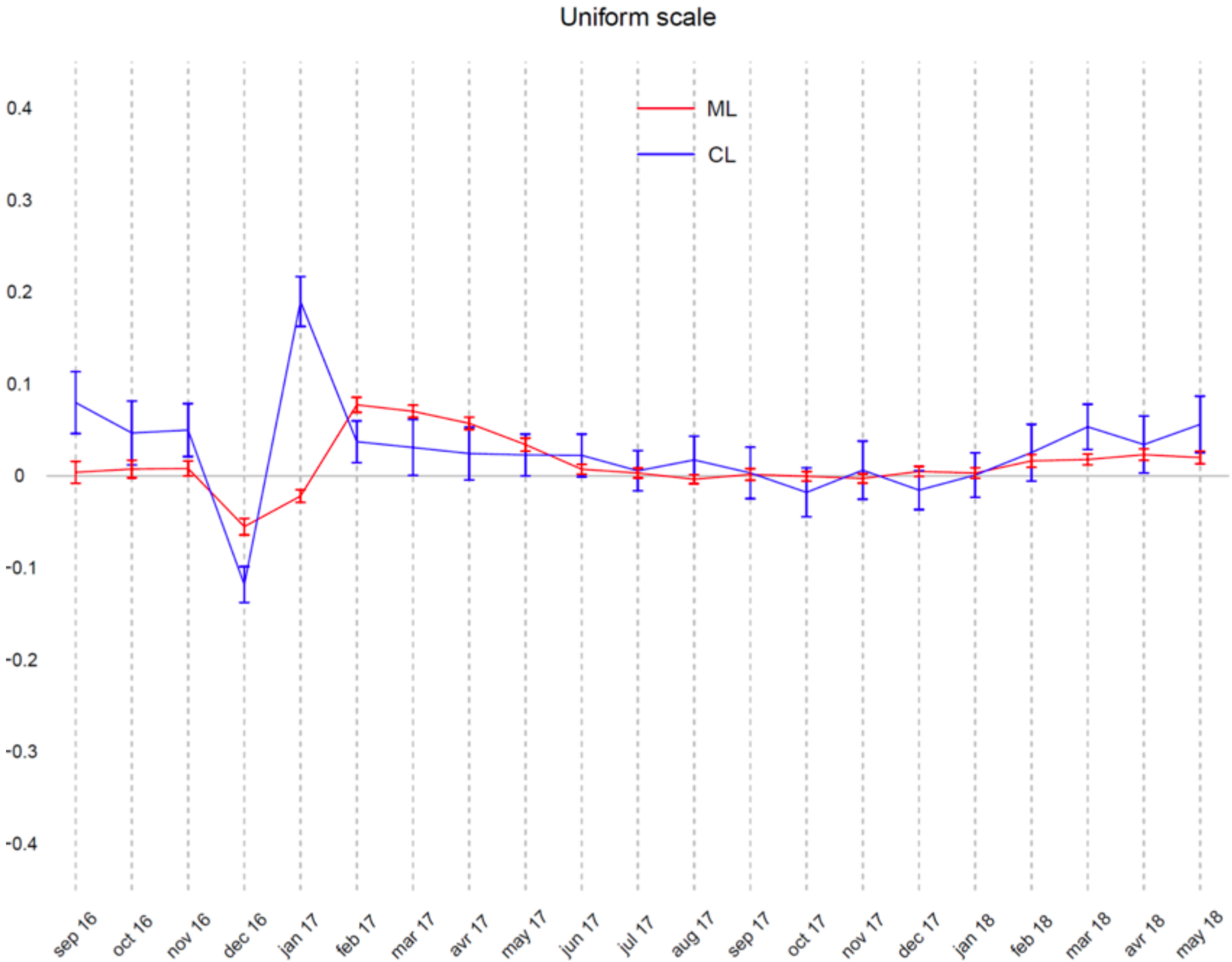


SHOCK ON CLAIM RATE FOR A SHORT PERIOD

Means of reserve predictions − positive shock on the claim rate



Means of relative errors for reserve predictions − positive shock on the claim rate



CONCLUSION

- We have proposed a **new non-parametric approach for individual claims reserving** using a machine learning algorithm known as Extra-Trees algorithm.
- Our model is **fully flexible** and **allow to consider (almost) any kind of feature information**.
- As a result we obtain I**BNR and RBNS claims reserves for individual policies** integrating all available relevant feature information.
- The method provides almost **unbiased estimators of the claims reserves with very small standard deviations** in our simulation study (almost four times smaller than the Mack chain-ladder standard deviation !).
- Our Machine Learning estimators are **more responsive to any changes in the development patterns of claims** including occurrence, reporting, cost modifications,... than the chain-ladder estimator based on aggregated loss data.