Dynamic Survival Modelling of Risk Credit Default

Ana María Patrón Piñerez
Supervised by: F. Javier Rubio Alvarez

September 4, 2025

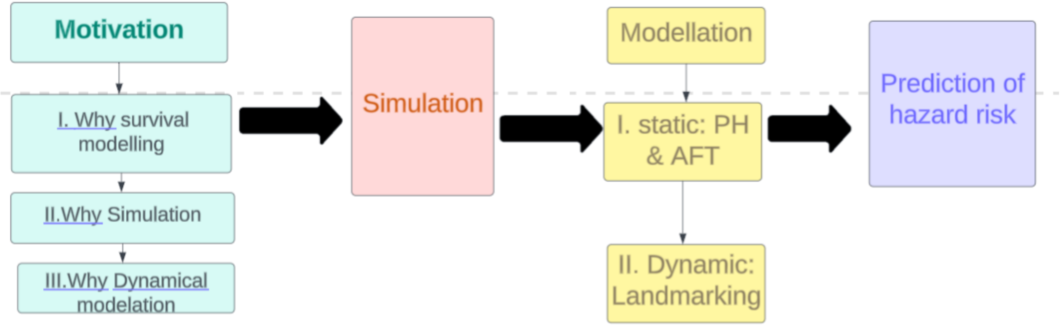
Mary Lister McCammon Summer Research Fellowship

Figure 1: Mary McCammon. First woman with a PhD in Mathematics at Imperial

Figure 2: Some Mary Lister interns 2025 and Heather Macbeth

What are we going to do?

Research Question: How can (bio)statistical models be used to study and predict UK housing credit defaults?



Survival Modelling

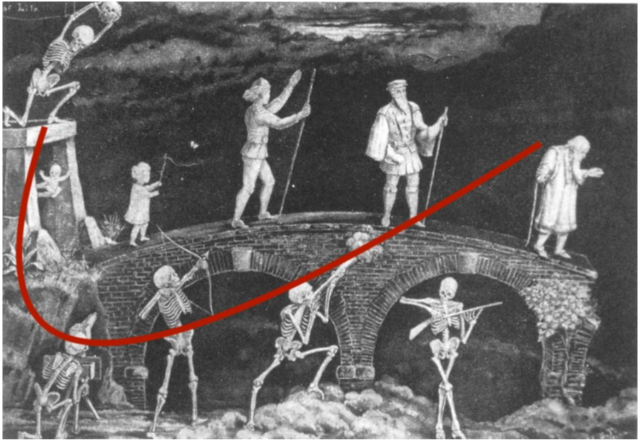
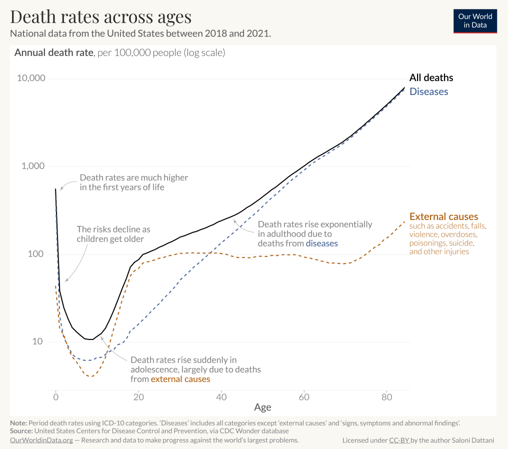


Figure 3: The Bridge of Life, Karl Pearson (1897)

I. Why Survival Modelling?

Hazard function: The **instantaneous risk** of an event happening at time t, given that the subject has survived up to t.

$$h(t) = \lim_{\Delta t \to 0} \frac{P[t \leq T < t + \Delta t \mid T \geq t]}{\Delta t}$$
$$= \frac{f(t)}{S(t)}$$



Survival Analogy: Biostatistics vs Finance

Moment 0	Birth or Diagnosis	→	Credit Granted
Event	Death	→	Credit Default
Interpretation	Probability of death at time t given it has been alive up to t	→	Probability of credit default at time t given it has not defaulted up to t

Survival Analogy: Biostatistics vs Finance

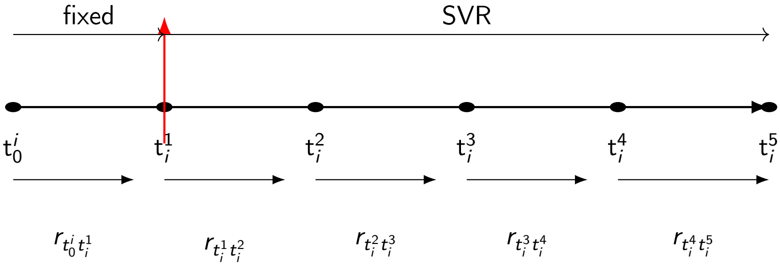
Moment 0	Birth or Diagnosis	Credit Granted
Event	Death	Credit Default
Interpretation	Probability of death at time t given it has been alive up to t	Probability of credit default at time t given it has not defaulted up to t

II. Why Simulated Data?

- Access to credit data is restricted due to UK GDPR and the Data Protection Act 2018.
- Case of **The Simulacrum**: a synthetic cancer dataset created by Health Data Insight CIC.
- Help researchers gain insights, write queries, and test analyses safely.
- Safe, free to use, and contains no real patient information.

III. Why Dynamical Survival Modelling?

Dynamic behaviour of the Mortgage sector



- **Fixed period:**
  Interest rate remains constant for 2, 3, or 5 years (depending on the contract).
- **Standard Variable Rate (SVR)**
  Interest rate usually changes following decisions by the Bank of England (every 6–8 weeks).

Generating times process timeline

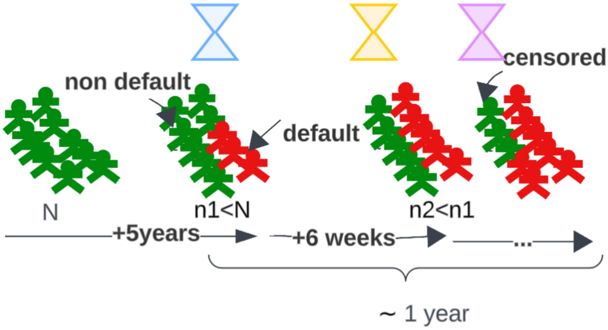
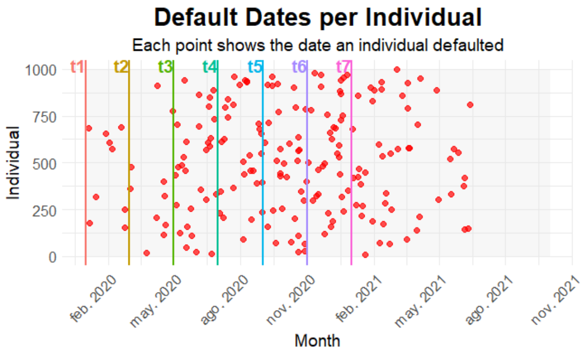


Figure 4: Simulation Diagram

Simulating default times



- Times are simulated as in Rubio, F. J. (2019):

$$t_i = F_0^{-1}\left(1 - \exp\left(\frac{\log(u_i)}{\exp(\boldsymbol{x_i}^\top \beta)}\right)\right)$$

where $u_i \sim \text{Unif}(0,1)$ at t_1 with rescaling at each iteration and baseline F_0 is a weibull.

- Monte Carlo simulations, truncated distributions for $\boldsymbol{x_i}$ (rate, inflation, LTV, age)

Static (and time-varying covariates)

- **Proportional Hazards (PH) Model:** Assumes the relative effect of covariates is **constant over time**

$$h(t \mid \mathbf{x}_i) = h_0(t)\exp(\mathbf{x}_i^\top \beta)$$

- **Accelerated Failure Time (AFT) Model:** Covariates **accelerate or delay the event** of default.

$$h(t \mid \mathbf{x}_i) = h_0\Big(t\exp(\mathbf{x}_i^\top \beta)\Big)\exp(\mathbf{x}_i^\top \beta)$$

Pros: Easy to use, interpretable, widely available in standard software (*R:* `Survival`,`GHSurv`, `flexsurv`).

Cons: Static (does not update risk over time); cannot handle time-varying covariates; may miscalibrate when individual characteristics change.

- **Landmarking**. For $t \geq t_L$

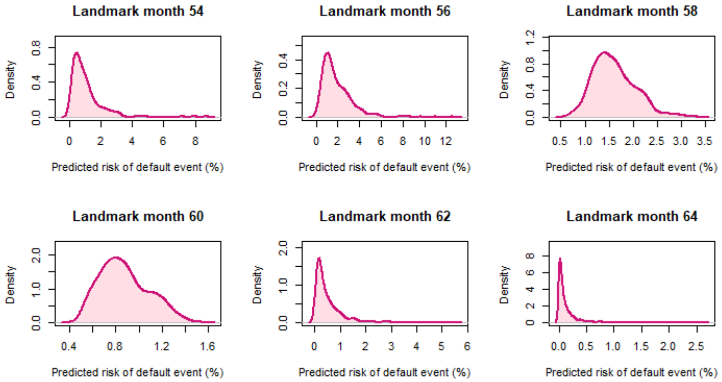
$$h(t \mid T > t_L, \mathbf{x}_i(t_L)) = h_0(t \mid t_L) \exp\left(\mathbf{x}_i(t_L)^\top \beta\right).$$

Pros: Captures changes in individuals' characteristics (dynamic).

Cons: Limited available software (*R:* `Landmarking*`); no unique way of updating information; more difficult interpretation.

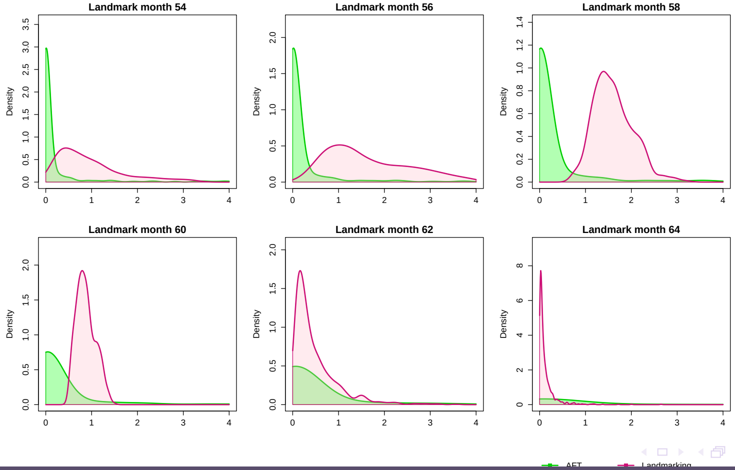
Predicted probabilities of credit default under dynamical setting

- Probabilities of credit default adjust dynamically



Comparison: Dynamical vs static model

- AFT does not recompute probabilities over time, whereas Landmarking does



Wraping up

- **Dynamic survival modelling for finance is a pioneering approach**, with limited literature currently available.
- **Limitations:** emulating real-world behavior remains challenging.
- **Future avenues:** improve data simulation and explore additional dynamic models for more realistic analyses.

References

- Rubio, F. J. (2019). *Simulating survival times from a General Hazard structure with a flexible baseline hazard*.
- Campbell, J. Y., & Cocco, J. F. (2015). *A Model of Mortgage Default*. The Journal of Finance, 70(4), 1495–1554.
- Barrott, I. (2022). *How to use the R package 'Landmarking'*.

In each case we state the model, estimation target, and the calculation of $S(t \mid X) = P(T > t \mid X) = 1 - F(t)$.
By definition, the conditional survival function is:

$$\begin{aligned}
S(t \mid x) &= \exp[-H(t \mid x)] \\
&= \exp[-\int_0^t h(s \mid x) ds] \\
&= \exp[-\int_0^t h_0(s \mid x) \exp(x^\top \beta) ds] \\
&= -H_0(t \mid x) \exp(x^\top \beta) \\
&= S_0(t) \exp(x^\top \beta)
\end{aligned}$$

Let $u_i \sim \text{Unif}(0,1)$ be a random variable. Then, for each iteration k:

$$S(t_1 \mid x) = u_i$$

$$S_0(t_1)\exp(x_i^\top \beta) = u_i$$

$$\exp(x_i^\top \beta)\log S_0(t_1) = \log u_i$$

$$S_0(t_1) = \exp\left(\frac{\log u_i}{\exp(x_i^\top \beta)}\right)$$

$$t_1 = F_0^{-1}\left(1 - \exp\left(\frac{\log u_i}{\exp(x_i^\top \beta)}\right)\right)$$

For the second iteration t_2, conditional on survival up to t_1:

$$S(t_2 \mid T > t_1, x) = \frac{S_0(t_2)}{S_0(t_1)}$$

$$u_2 \sim \text{Unif}(0, S_0(t_1))$$

$$S_0(t_2) = \exp\left(\frac{\log u_2}{\exp(x_i^\top \beta)}\right)$$

$$t_2 = F_0^{-1}\left(1 - \exp\left(\frac{\log u_2}{\exp(x_i^\top \beta)}\right)\right)$$

For the third iteration t_3, conditional on survival up to t_2:

$$S(t_3 \mid T > t_2, x) = \frac{S_0(t_3)}{S_0(t_2)}$$

$$u_3 \sim \mathrm{Unif}(0, S_0(t_2))$$

$$S_0(t_3) = \exp\left(\frac{\log u_3}{\exp(x_i^\top \beta)}\right)$$

$$t_3 = F_0^{-1}\left(1 - \exp\left(\frac{\log u_3}{\exp(x_i^\top \beta)}\right)\right)$$

In general, for the k-th iteration:

$$S(t_k \mid T > t_{k-1}, x) = \frac{S_0(t_k)}{S_0(t_{k-1})}$$

$$u_k \sim \mathrm{Unif}(0, S_0(t_{k-1}))$$

$$S_0(t_k) = \exp\left(\frac{\log u_k}{\exp(x_i^\top \beta)}\right)$$

$$t_k = F_0^{-1}\left(1 - \exp\left(\frac{\log u_k}{\exp(x_i^\top \beta)}\right)\right)$$

For the third iteration t_3, conditional on survival up to t_2:
Suppose we have landmarking intervals $[t_k, t_{k+1})$ and a Weibull survival function:

$$S_0(t) = \exp\left(- \left(\frac{t}{\lambda}\right)^k \right)$$

We want the probability of death in each interval to be approximatly 0.2%, i.e.,

$$\Pr(t_k \leq T < t_{k+1}) = 0.02$$

The probability of the event given survival up to t_k is:

$$\Pr(t_k \leq T < t_{k+1} \mid T \geq t_k) = 1 - \frac{S_0(t_{k+1})}{S_0(t_k)}$$

Optimization of mortality ratio II

Therefore, we impose the condition:

$$1 - \frac{S_0(t_{k+1})}{S_0(t_k)} = 0.02 \quad \Longrightarrow \quad \frac{S_0(t_{k+1})}{S_0(t_k)} = 0.98$$

Substituting the Weibull survival function:

$$\frac{\exp\Big(-(t_{k+1}/\lambda)^k\Big)}{\exp\Big(-(t_k/\lambda)^k\Big)} = 0.8 \quad \Longrightarrow \quad \exp\Big(-(t_{k+1}/\lambda)^k + (t_k/\lambda)^k\Big) = 0.98$$

Taking the natural logarithm:

$$(t_{k+1}/\lambda)^k - (t_k/\lambda)^k = -\log(0.98)$$

Finally, solving for λ for that interval:

$$\lambda = \left[\frac{(t_{k+1})^k - (t_k)^k}{-\log(0.8)} \right]^{1/k}$$

This ensures that aproximately 0.2% of the population alive at the start of the interval dies during that interval.