شرح الورقة بالمصري — Foundation or Formula?

يعني إيه "Formula" و يعني إيه "Foundation"؟

عنوان الورقة بيحطّ كلمتين قدام بعض. خلينا نفهمهم الأول، لأن الورقة كلها مبنية على المقارنة بينهم.

1) الـ Formula — "الصيغة"، الطريقة القديمة

الـ **Formula** معناها **معادلة رياضية** إنت بتفترضها وبعدين بتظبّط أرقامها على بياناتك.

مثال: `ARIMA` بتقول لك إن قيمة النهارده = معادلة فيها قيم الأيام اللي فاتت + الأخطاء اللي فاتت. المعادلة نفسها **إنت** اللي بتختارها، والكمبيوتر بيقعد يقدّر الأرقام (الـ parameters) اللي جواها من بياناتك إنت لوحدك.

خصائصها:

- **صغيرة جداً** — 3 أو 4 أرقام بس.
- **بتتعلّم من سلسلتك إنت بس** — معندهاش أي معرفة سابقة بالدنيا.
- **مفهومة** — تقدر تبصّ على الأرقام وتفهم الموديل بيقول إيه وليه.
- **قديمة ومجرَّبة** — بقالها 50 سنة في الكتب والصناعة.

من ديه: `Seasonal Naive`، `Theta`، `ETS`، `ARIMA`.

2) الـ Foundation — "الأساس"، الطريقة الجديدة

الـ **Foundation Model** معناها **شبكة عصبية ضخمة اتدرّبت مرة واحدة على ملايين السلاسل الزمنية من الدنيا كلها**، وبعدين بتديها **سلسلتك إنت** فتتنبأ ليك **من غير ما تدرّبها من أصله**. ودي اللي بيسمّوها **zero-shot**.

`TimesFM-3` من ديه: **330 مليون رقم** جواه، متدرّب على **أكتر من تريليون نقطة زمنية**.

خصائصها:

- **ضخمة** — 330 مليون رقم مقابل 3.
- **بتعرف حاجات من برّه سلسلتك** — شافت ملايين السلاسل قبل كده.
- **مقفولة** — مش هتعرف تبصّ جواها وتفهم قرارها.
- **جديدة** — الحاجات دي طلعت في آخر سنتين.

الكلمة المفتاحية: مصدر المعرفة

الفرق الحقيقي بينهم مش "قديم وجديد" ولا "صغير وكبير" — الفرق هو **المعرفة جاية منين**:

	Formula	Foundation
بتعرف من فين؟	من **سلسلتك إنت** بس	من **ملايين السلاسل** اللي اتدرّبت عليها
محتاجة إيه؟	بيانات كفاية عشان تقدّر معادلتها	مفيش — بتشتغل على طول
بتنفع إمتى؟	لما يبقى عندك بيانات كتير وشكلها واضح	لما بياناتك **قليلة** أو شكلها **غريب**

تشبيه: الـ **Formula** زي **دكتور خرّيج لسه** بس معاه كتاب مظبوط: لو الحالة موصوفة في الكتاب بالظبط، هيبقى ممتاز. والـ **Foundation** زي **دكتور شاف مليون حالة**: مش معاه كتاب، بس عنده **حدس** قوي — ولو الحالة نادرة هيبقى أحسن منه بكتير.

وعشان كده العنوان سؤال

"Foundation or Formula?" يعني: **أستخدم الوحش الضخم اللي شاف الدنيا، ولا المعادلة الصغيرة المفهومة؟**

والورقة **مش** بتجاوب "ده أحسن من ده". بتجاوب: **إمتى ده وإمتى ده** — وبأرقام متختبرة إحصائياً مش بآراء.

من الآخر: `Formula` = معادلة صغيرة بتتعلّم من بياناتك إنت. `Foundation` = شبكة ضخمة اتعلّمت من بيانات الدنيا. والسؤال هو مين فيهم ينفعك، **إمتى**.

---

0. The Big Idea in One Breath — الفكرة الكبيرة في نَفَس واحد

قبل ما ندخل في أي تفصيلة، خد الصورة كلها في دقيقة:

Google نزّلت موديل اسمه `TimesFM-3` بيتنبأ بالسلاسل الزمنية (time series) من غير ما تدرّبه على بياناتك خالص — بيسمّوا ده **zero-shot**. السؤال العملي: أنا كـ محلل، لو قدامي سلسلة زمنية، أستخدم الموديل الجبار ده ولا أستخدم `ARIMA` القديمة الكويسة؟

الورقة دي بتجاوب على السؤال ده. وأهم حاجة فيها إننا **مخترعناش طريقة جديدة** — دي دراسة مقارنة (comparative study)، وقلناها بصراحة في الورقة. اللي يخليها تستاهل النشر حاجة تانية: **إحنا مولّدنا البيانات بنفسنا**، والسبب هنشرحه بعد شوية وهو قلب الموضوع كله.

النتيجة باختصار: **مفيش طرف كسب**. لكن الطريقة اللي "مكسبش" بيها كل طرف هي الحكاية الحلوة.

من الآخر: الورقة مش بتقول "الـ AI كسب" ولا "الكلاسيكي أحسن" — بتقول **إمتى** ده وإمتى ده، وبأرقام متختبرة إحصائياً.

---

1. What Is TimesFM-3? — إيه هو الـ TimesFM-3 أصلاً؟

في **31 أغسطس 2026** نزلت Google الجيل التالت من موديلها `TimesFM` . خصائصه:

- **330 مليون parameter** (بالظبط 330,710,976 — إحنا عدّيناهم من الـ checkpoint نفسه مش نقلاً عن حد).
- متدرّب على **أكتر من تريليون نقطة زمنية** (time points)، بيانات حقيقية **و** بيانات صناعية (synthetic).

- أول واحد في العيلة متدرّب أصلاً على الـ **multivariate** — يعني ياخد أكتر من سلسلة مع بعض، ويقدر ياخد `covariates` (متغيرات مساعدة) سواء من الماضي أو معروفة للمستقبل.
- بيستخدم `attention` بالتبادل: مرة على المحور الزمني (causal-temporal) ومرة على محور المتغيرات (full-variate).
- بيطلّع **الأفق كله في مرة واحدة** (non-autoregressive) — يعني مش بيتنبأ بخطوة ويرجع يدخّلها عشان يتنبأ باللي بعدها.
- بيطلّع **9 quantiles** (من `q0.1` لـ `q0.9`) — يعني مش رقم واحد، دا توزيع.
- طول الـ context بيوصل **16k** نقطة.

تشبيه: تخيل دكتور اتخرّج ومشتغلش في مستشفى واحدة بس — لأ، ده اتفرّج على مليون حالة من مستشفيات الدنيا كلها. لما ييجي يشوف مريضك، هو معندوش ملف مريضك بس عنده **حدس** اتبنى من كل اللي شافه. ده بالظبط معنى `zero-shot`: مشافش سلسلتك، بس شاف حاجات شبهها كتير أوي.

من الآخر: `TimesFM-3` = موديل ضخم بيتنبأ من غير تدريب، وبيديك توزيع مش رقم.

---

2. Why the Published Evidence Is Not Enough — ليه الأدلة المنشورة مش كفاية؟

هنا لب المشكلة، وده اللي بنى عليه البحث كله.

لما شركة تنزّل موديل، بتقولك "بصّ، كسبت على الـ benchmarks الفلانية". المشكلة إن الـ benchmarks دي بيانات **عامة ومنشورة على الإنترنت**، والموديل اتدرّب على... الإنترنت.

باحثين اسمهم Meyer وزمايله (2025) عملوا حاجة مهمة: تتبّعوا **نسب البيانات** لـ 22 موديل foundation model للسلاسل الزمنية. لقوا إن من **401 dataset** استُخدمت في تقييمهم، **6% بس** هي اللي **مظهرتش** في corpus التدريب بتاع أي موديل منهم.

يعني إيه؟ يعني 94% من التقييمات دي، الموديل يمكن يكون **شايف الإجابة قبل الامتحان**.

وأكتر من كده: لقوا إن تداخل بنسبة **0.1% بس** بين بيانات التدريب وبيانات الاختبار ممكن يحرّك الـ MAPE بمقدار **8 لـ 29 نقطة مئوية**. ودي مش نسبة صغيرة، دي فرق بين موديل ممتاز وموديل فاشل.

وكمان — والنقطة دي ذكية — التسريب (leakage) مش لازم يكون **مباشر**. لو الموديل اتدرّب على سلسلة تانية **متداخلة زمنياً** ومرتبطة سببياً بسلسلتك، هو كده أخد ميزة على سلسلتك حتى لو عمره ما شافها.

تشبيه: طالب ذاكر من بنك أسئلة، والامتحان جه 94% منه من بنك الأسئلة ده. درجته العالية معناها إيه بالظبط؟ مش عارفين. ودي بالظبط مشكلة الـ benchmarks.

فيه كمان انتقاد تاني اتقال كتير على الأوراق دي: بيقارنوا الموديل الجبار بـ baselines كلاسيكية **مفكوكة وغير متظبطة** — يعني ياخدوا `ARIMA` لوحدها بدل ما ياخدوا الـ **combination** اللي بتكسب في مسابقات التنبؤ زي `M4`. ده زي ما تقارن فريقك بلاعب واحد من الفريق التاني.

من الآخر: أي رقم من benchmark عام = قياس **مجهول المصدر**، ولازم نلاقي طريقة تانية.

---

3. The Idea That Closes Both Objections — الفكرة اللي بتقفل الاعتراضين مع بعض

الحل اللي الورقة اختارته: **إحنا نولّد البيانات بنفسنا**.

لما نولّد سلسلة من عملية معروفة (Data-Generating Process — واختصارها `DGP`)، بيحصل **حاجتين في نفس اللحظة**، وده اللي بيدي المقارنة قوتها:

1. **الموديل قطعاً مشافش السلسلة دي** — مش "على الأغلب"، لأ: السلسلة دي **معملتش أصلاً** غير لما شغّلنا التجربة. فمستحيل تكون في corpus التدريب.
2. **الموديل الكلاسيكي بيبقى "صح بالتعريف"** (correctly specified by construction) — لو البيانات فعلاً `ARIMA(1,1,1) with drift`، يبقى `AutoARIMA` مش تقريب للحقيقة، دي **هي الحقيقة** نفسها، وفاضل بس إنه يقدّر الـ parameters.

النقطة التانية دي هي السر. عشان كده السؤال بقى حاد ومحدد:

- الموديل بيخسر قد إيه لما يلعب على **أرض الكلاسيكي** (يعني الشكل الرياضي صح 100%)؟
- وبيكسب قد إيه لما **مفيش شكل كلاسيكي صح** (كسر هيكلي، لا خطية، طلب متقطع، تقلّب)؟

⚠️ تفرقة مهمة جداً — لازم تكون أمينة

فيه حاجة لازم نقولها بصراحة، وقلناها في الورقة:

- اللي التصميم بتاعنا **بيمنعه** هو التلوث على مستوى **التحقق الفردي** (the realisation) — السلسلة دي بالذات مكنتش موجودة.
- اللي التصميم **مش** بيمنعه هو الألفة على مستوى **العيلة** (the family). يعني `TimesFM-3` اتدرّب على بيانات **صناعية** كمان، والبيانات الصناعية للـ time-series models عادةً بتتولّد من نفس العائلات دي بالظبط (`ARMA`, `AR`, seasonal..). فهو **يعرف الشكل العام** حتى لو مشافش النسخة دي.

عشان كده عنوان الورقة بيقول **"A Simulation-Based Comparison"** — بيوصف الطريقة (محاكاة) من غير ما يدّعي إن الموديل مايعرفش نوع البيانات، لأن ده ادعاء أوسع من اللي التصميم بيثبته.

من الآخر: إحنا مولّدنا البيانات عشان نضمن حاجتين: الموديل مشافهاش، والكلاسيكي صح بالتعريف. بس مدّعيناش إن الموديل مش عارف نوع البيانات — دي أمانة علمية مش تواضع.

---

4. Design of the Simulation — تصميم دراسة المحاكاة

4.1 The Nine DGPs — التسع عمليات المولّدة للبيانات

اخترنا 9 عمليات بحيث تغطي المدى من "الكلاسيكي صح تماماً" لـ "مفيش كلاسيكي صح":

الرمز	العملية	الشرح بالمصري	مين "صح" فيها؟
D1	AR(1), φ=0.7	القيمة النهارده = 0.7 × امبارح + عشوائية. أبسط حاجة ممكنة.	ARIMA
D2	ARMA(1,1)	زي فوق + بتفتكر **الخطأ** بتاع امبارح كمان.	ARIMA
D3	ARIMA(1,1,1) + drift	فيها **اتجاه** (trend) — بتمشي لفوق بمعدل ثابت + عشوائية.	ARIMA
D4	SARIMA(1,0,0)(1,1,0)[12]	فيها **موسمية** قوية كل 12 شهر (زي مبيعات رمضان).	ARIMA
D5	ETS(A,A,A)	مستوى + اتجاه + موسمية، بس بصيغة الـ state-space.	ETS
D6	Local level + كسر هيكلي	ماشية عادي وفجأة **تقفز** قفزة (زي زيادة سعر).	ولا واحد
D7	Logistic growth	نمو بيتشبّع (S-curve) — بيطلع بسرعة وبعدين يقف.	ولا واحد
D8	Intermittent demand	طلب **متقطّع**: 70% أصفار، وبين الحين والتاني طلبية.	ولا واحد
D9	AR(1) + GARCH(1,1)	المتوسط صح، بس **التذبذب نفسه بيتغير** (فترات هدوء وفترات جنون).	ARIMA في المتوسط بس

ليه اخترنا D6 بالكسر جوه الـ context مش في المستقبل؟ دي نقطة تصميم مهمة جداً. لو حطينا الكسر في الفترة اللي بنتنبأ بيها، **مفيش طريقة في الدنيا** تقدر تتنبأ بيه — فالتجربة كانت هتقيس بس "مين أكتر واحد محافظ ومش بيتحرك". لكن لما نحطه **جوه البيانات اللي الموديل شافها** (عند 0.7 × n)، السؤال بيبقى عادل ومحترم: "إنت شفت المستوى اتغيّر — هتتنبأ من المستوى الجديد ولا هترجع للقديم؟"

ليه D7 نقطة الانقلاب عند 0.6 من السلسلة؟ عشان الفترة اللي بنتنبأ بيها تقع في **الجزء اللي بيتشبّع**. أي طريقة بتركّب خط مستقيم (drift) هتـ **تعدّي فوق** (overshoot). وده بالظبط اللي عايزين نكشفه.

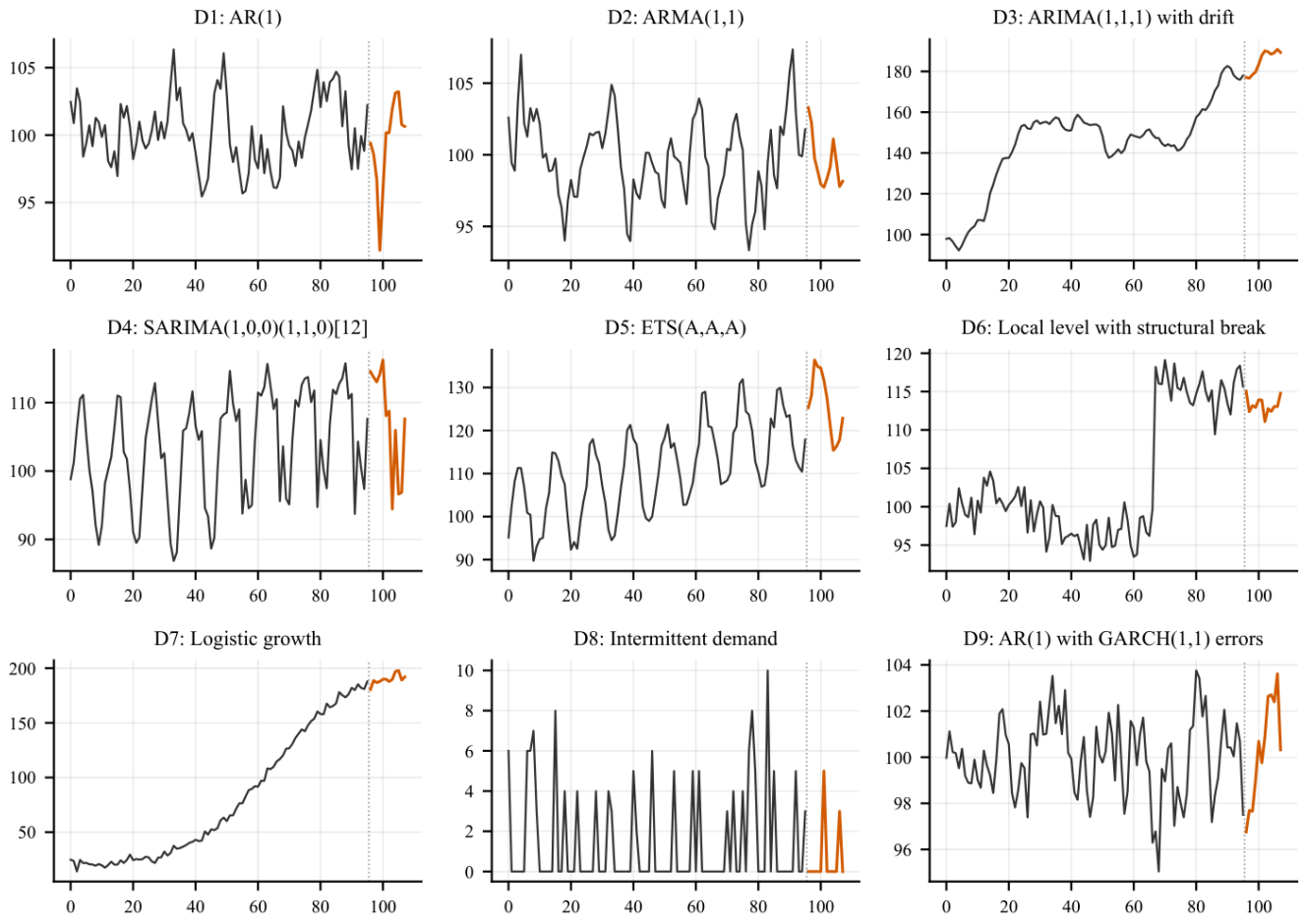
4.2 Lengths, Replications, Horizon — الأطوال والتكرارات والأفق

- **أطوال السلاسل:** `n ∈ {24, 48, 96, 200}` . الـ 24 دي قاسية بشكل مقصود: مع موسمية 12، يبقى عندك **دورتين موسميتين بس**. وهنشوف بعدين إن ده بيفرق حاجة كانت الناس بتخلط بينها.
- **التكرارات:** 200 تكرار لكل خلية (DGP × طول).
- **إجمالي السلاسل:** 9 × 4 × 200 = **7,200 سلسلة**.
- **الأفق (horizon):** `H = 12` . بنولّد `n + 12` نقطة، أول `n` بندّيها للموديل، وآخر 12 بنخبّيها ونقارن بيها.
- **الـ Seeding:** كل سلسلة ليها seed محسوبة من هويّتها:

`رقم الطول + رقم التكرار × 1,000 + DGP رقم الـ × 1,000,000 = seed` . يعني السلسلة بتتعاد **بالظبط** من غير ما يفرق عدد الأنوية اللي شغّالة ولا ترتيب التنفيذ.

من الآخر: 9 عمليات × 4 أطوال × 200 تكرار = 7,200 سلسلة، كل واحدة ليها seed ثابتة، فأي حد يقدر يعيد نفس الأرقام بالظبط...

Figure 1: One realisation of each data-generating process (n = 96; held-out horizon in colour)



شكل 1 — عيّنة واحدة من كل عملية من التسعة (عند `n = 96`). الجزء الرمادي هو البيانات اللي الموديل بيشوفها، والجزء الملوّن في الآخر هو الـ 12 نقطة المخبّية اللي بنقيس عليها. بصّ على `D6`: الكسر الهيكلي واضح عند النقطة 67 تقريباً — يعني `0.7 × 96` بالظبط زي ما صمّمناه. وبصّ على `D8`: معظمها أصفار، وده اللي هيبوّظ الـ `MAPE` بعدين.

5. The Methods Compared — الطرق اللي اتقارنت

مفيش أي طريقة اتظبطت بالإيد. كل طريقة كلاسيكية شغّالة بالاختيار الأوتوماتيكي بتاعها — لأن ده الوضع اللي المحلل الحقيقي بيقابله.

الطريقة	الشرح بالمصري
`TimesFM3`	الموديل بتاع Google، zero-shot، من غير تدريب ولا تظبيط.
`SeasonalNaive`	أغبى طريقة ممكنة: "النهارده هيبقى زي نفس اليوم السنة اللي فاتت".
`Theta`	الطريقة اللي كسبت مسابقة `M3` — بتفكّك السلسلة وتعيد تركيبها.
`AutoETS`	Exponential smoothing بيختار الشكل أوتوماتيك (Hyndman-Khandakar).
`AutoARIMA`	ARIMA بيدوّر على الـ orders أوتوماتيك.
`Combination`	**متوسط** الـ 3 دول: AutoARIMA. + AutoETS + Theta

ليه الـ Combination إجباري مش اختياري؟

عشان أشهر انتقاد بيتقال على أوراق الـ foundation models هو إنهم بيقارنوا نفسهم بطرق كلاسيكية **مفكوكة**. الـ state of the art الكلاسيكي هو **الدمج** مش الطريقة الواحدة. لو شلناه، يبقى إحنا بنينا النتيجة اللي عايزينها **جوه التصميم** نفسه — وده غش منهجي.

⚠️ ملاحظة أمانة: عدم تماثل في المعلومات

فيه حاجة قلناها في الورقة بصراحة لأنها **في صالح الكلاسيكي**: كل الطرق الكلاسيكية بندّيها **الفترة الموسمية الحقيقية** (m = 12) كمعلومة جاهزة، بينما `TimesFM-3` بياخد السلسلة الخام من غير ما نقوله إن فيها موسمية أصلاً ولا طولها كام. يعني الكلاسيكي واخد مساعدة.

من الآخر: 6 طرق، الكلاسيكي واخد مساعدة (الموسمية جاهزة)، والـ Combination موجود عشان المقارنة تبقى شريفة.

---

6. The Metrics — المقاييس، وإزاي بنقيس "الغلط"

6.1 Point Metrics — مقاييس النقطة الواحدة

`MASE` **(المقياس الأساسي):** الخطأ المطلق مقسوم على خطأ الـ seasonal naive **داخل العينة**.

$$\text{MASE} = \frac{\frac{1}{h}\sum_{i=1}^{h}|y_i - \hat{y}_i|}{\frac{1}{n-m}\sum_{t=m+1}^{n}|y_t - y_{t-m}|}$$

تفكيك المعادلة رمز رمز:

- y_i — القيمة **الحقيقية** في الخطوة رقم i من الفترة اللي بنتنبأ بيها.
- $\hat{y}_i$ — القيمة **المتوقعة** من الموديل لنفس الخطوة. (القبعة فوق الحرف معناها "تقدير").
- h — **طول الأفق**، يعني كام خطوة بنتنبأ بيها. عندنا $h = 12$.
- n — **طول السلسلة المعروفة** اللي الموديل شافها.
- m — **الفترة الموسمية**. عندنا $m = 12$ للسلاسل الموسمية، و $m = 1$ لغيرها.

والكسر نفسه:

- **البسط** — متوسط الخطأ المطلق **بتاعك** في الفترة المتنبأ بيها. يعني: بتغلط بكام في المتوسط؟
- **المقام** — متوسط خطأ **"أغبى طريقة"** (الـ seasonal naive: "النهارده زي نفس اليوم في الدورة اللي فاتت")، محسوب على البيانات **اللي إنت شايفها** مش على المستقبل.

وإزاي تقرا الرقم الناتج؟

- $\text{MASE} = 1$ ← إنت **بالظبط زي** الطريقة الغبية.
- $\text{MASE} < 1$ ← إنت **أحسن** منها. يعني `0.68` معناها خطأك 68% من خطأها.
- $\text{MASE} > 1$ ← إنت **أوحش** من الطريقة الغبية! وده بيحصل فعلاً — شوف `AutoETS` في القسم 8.3.

ليه القسمة دي مهمة؟ عشان تقدر تجمع نتايج من سلاسل مقاييسها مختلفة تماماً — سلسلة أرقامها بالملايين وسلسلة أرقامها بالعشرات — من غير ما الكبيرة تبلع الصغيرة.

`RMSSE` : نفس الفكرة بالظبط بس بالتربيع بدل المطلق (زي ما في مسابقة **`sMAPE`**). `M5` : نسبة مئوية متماثلة، مدى من 0 لـ 200.

6.2 Why MAPE Is Not the Headline — ليه الـ MAPE مش المقياس الأساسي

الـ `MAPE` هو أشهر مقياس عند الناس، ومع ذلك حطّيناه في **الملحق** بس. ليه؟

1. **غير معرّف عند الصفر** — القسمة على صفر. وفي `D8` (الطلب المتقطع) 70% من القيم أصفار!
2. **غير متماثل** — بيعاقب المبالغة في التقدير أكتر من التقليل. فهو **بيحابي** الطرق المتحفظة.
3. **بيقلب الترتيب** — وده مش كلام نظري، ده حصل في بياناتنا فعلاً.

الرقم اللي بيثبت الكلام ده: الـ MAPE طلع **غير معرّف في 12,906 من 129,600** تقييم — يعني **9.96%** — وكلهم **بالظبط** في `D8` .

يعني لو كنا اعتمدنا على الـ MAPE، كنا هنضطر **نشيل النظام اللي فيه أقوى نتيجة عندنا** من المقارنة خالص، أو نخترع اتفاقية للأصفار من دماغنا.

من الآخر: الـ MAPE مقياس مشهور لكنه بيكدب لما فيه أصفار — و `D8` كلها أصفار، فمينفعش يكون الأساس.

6.3 Probabilistic Metrics — مقاييس التوزيع مش النقطة

الموديلات دي مش بتديك رقم واحد، بتديك **توزيع**. فلازم نقيس التوزيع نفسه:

- **`Pinball loss`** **(خسارة الـ quantile):** بنحسبها على الـ 9 deciles كلها، ومقسومة على نفس مقام الـ MASE عشان تبقى قابلة للمقارنة. دي **proper scoring rule** يعني مقياس شريف رياضياً.
- **`Coverage`** : الفترة اللي قلت عليها 80% — هل الحقيقة وقعت جواها 80% من المرات فعلاً؟
- **`Width`** : عرض الفترة. ودي مهمة أوي — هنشوف ليه كمان شوية.

⚠️ ليه 60% و 80% ومفيش 90%؟

دي نقطة تصميم مقصودة: `TimesFM-3` بيطلّع **الـ deciles بس** (من `q0.1` لـ `q0.9`). فمفيش عندك `q0.05` ولا `q0.95` عشان تعمل فترة 90%. لو قارنّا فترة 90% كلاسيكية بفترة معمولة بالاستقراء (extrapolation) من الموديل، تبقى

مقارنة **مش عادلة**. فأخدنا أوسع فترة **الاتنين** يقدروا يعبّروا عنها: `[q0.1, q0.9]` .

من الآخر: بنقيس النقطة (MASE) والتوزيع (pinball) والفترة (width + coverage) — التلاتة، مش واحد.

---

7. The Statistical Tests — الاختبارات الإحصائية

مش كفاية نقول "المتوسط ده أقل من ده". لازم **نختبر**.

7.1 Paired Wilcoxon — الاختبار الأساسي

جوه كل خلية (DGP × طول × أفق)، بنقارن `TimesFM-3` بكل طريقة كلاسيكية باختبار **Wilcoxon signed-rank المزدوج** على فروق الـ MASE بين الـ 200 تكرار.

ليه Wilcoxon مش t-test؟ لأنه **مش بيفترض التوزيع الطبيعي**، وتوزيعات الخطأ في التنبؤ بتبقى ملتوية (skewed) بشكل بشع. بس برضه بنحسب الـ `t-test` كـ **sensitivity check**.

7.2 Why NOT Diebold–Mariano — ليه رفضنا اختبار Diebold–Mariano

دي نقطة منهجية مهمة، والأدب المنشور بيستخدم `DM` بشكل شبه تلقائي.

الـ `DM` **متصمم** لسلسلة فروق خسارة **مترابطة زمنياً** (autocorrelated) جاية من **سلسلة واحدة** عبر الزمن. عندنا في المحاكاة، الـ 200 تكرار دول **سحوبات مستقلة** من نفس العملية — مفيش ترتيب زمني بينهم أصلاً! فلو استخدمنا DM هنبقى بنوصف بنية اعتماد **مش موجودة**.

وفي بيانات `M4` الحقيقية، التصميم بتاعنا فيه **3 origins بس** لكل سلسلة، يعني الإحصاءة هتبقى بـ **درجتين حرية** — يعني مفيش قوة إحصائية خالص.

⚠️ **وحصل غلط حقيقي هنا اكتشفناه بالمراجعة:** في نسخة أولى، الكود حسب DM على **مقطع عرضي** (cross-section) من 1000 متوسط مرتبين **أبجدياً** بحسب اسم السلسلة! يعني التصحيح كان بيجمع autocovariances عبر سلاسل **مالهاش علاقة ببعض** وبترتيب **عشوائي تماماً**. الدليل القاطع: الإحصاءة طلعت **-13.97** بالترتيب ده، وبين **-18 و -23** لما نخلط الترتيب عشوائياً. حاجة بتتغير 30–70% لما تخلط فهرس عشوائي دي **مش اختبار**.

فقررنا **نمسحها خالص** مش نصلّحها، لأن مفيش نسخة منها التصميم ده بيسمح بيها.

7.3 Multiplicity — التصحيح للمقارنات المتعددة

إحنا بنعمل 180 مقارنة. لو كل واحدة عند 5%، يبقى متوقع نلاقي 9 نتايج "معنوية" **بالصدفة البحتة**. فبنستخدم `Benjamini–Hochberg` للتحكم في الـ False Discovery Rate عند 0.05، وبنبلّغ الـ p الخام والمصحّح.

7.4 Effect Sizes — أحجام الأثر

مع كل اختبار بنبلّغ: - **`Hodges–Lehmann`** : تقدير للفرق الوسيط. - **نسبة الـ MASE الوسيطة:** عشان تشوف الفرق **عملياً** يعني إيه.

ليه؟ عشان فرق ممكن يبقى **معنوي إحصائياً** وهو **تافه عملياً**. مع 200 تكرار، فرق 0.2% ممكن يطلع معنوي. لازم القارئ يشوف ده.

من الآخر: بنختبر كل فرق، بنصحّح للمقارنات المتعددة، وبنقول حجم الفرق مش بس هل هو معنوي.

---

8. The Results — النتائج

التصميم طلّع 7,200 سلسلة، ومع 6 طرق و 3 شرائح أفق = **129,600 تقييم**.

8.1 A Split Decision, Not a Rout — قرار منقسم مش كاسح

لو عدّينا **مين كسب في كل خلية** من الـ 36 خلية (DGP × طول):

الطريقة	عدد الخلايا اللي كسبتها
`TimesFM3`	**16**
`AutoARIMA`	10
`SeasonalNaive`	4
`Theta`	2
`Combination`	2
`AutoETS`	2

يعني الكلاسيكي كسب 20 والموديل كسب 16. القراءة السطحية: "الكلاسيكي كسب".

بس دي قراءة ناقصة، وهنا بيت القصيد: الـ 20 دول **متوزعين على 5 طرق مختلفة**! يعني عمود الكلاسيكي هو عمود **فايزين مختلفين**، وإنت **معندكش طريقة تعرف مقدماً** مين منهم هيكسب على سلسلتك من غير ما تعمل الشغل اللي الـ foundation model المفروض يوفره عليك.

ولما نختبر ضد **أحسن** طريقة كلاسيكية في كل خلية: **18 من 36 فرق** نجوا من تصحيح BH — **7 لصالح TimesFM-3** و **11 لصالح الكلاسيكي**، و**18 خلية تعادل**.

8.2 Who Actually Holds Their Ground — مين بيصمد قدامه فعلاً؟

خلينا نبص على الـ 180 مقارنة الثنائية. منها **132 معنوية**: **113 لصالح TimesFM-3** و **19 لصالح الكلاسيكي**. التفصيل حسب الخصم:

الخصم	مقارنات معنوية	كسب TimesFM-3	كسب الخصم
`AutoARIMA`	20	**11**	**9**
`SeasonalNaive`	26	22	4
`AutoETS`	27	25	2
`Theta`	29	28	1
`Combination`	30	27	3

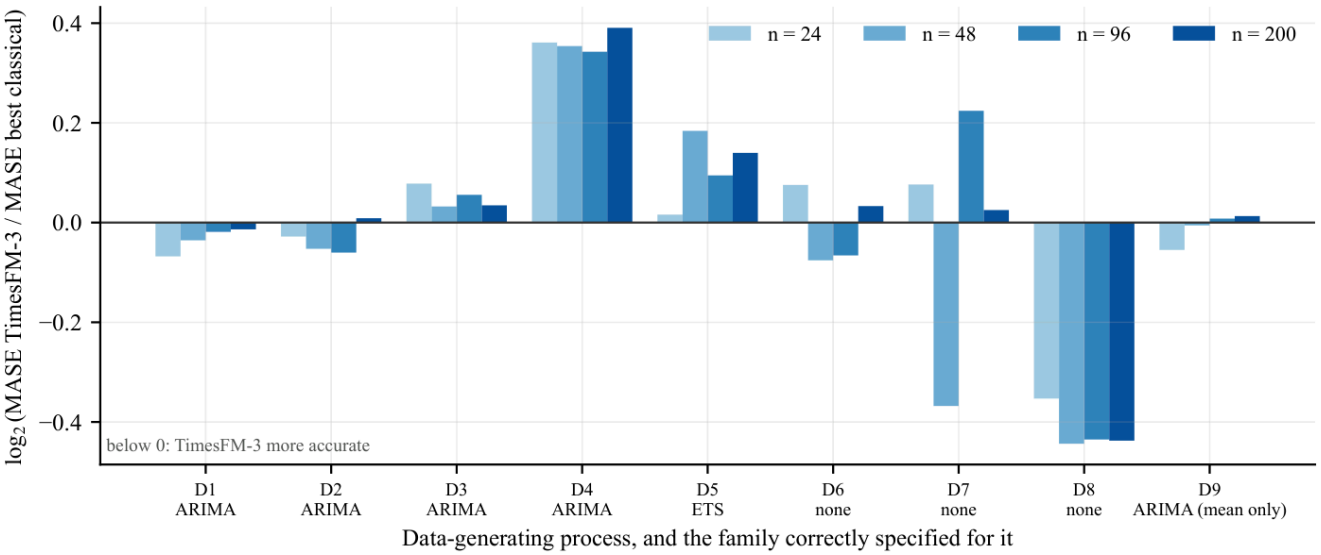
الجدول ده هو رسالة الورقة المركزية. بصّ على السطر الأول: `AutoARIMA` هو **الوحيد** اللي بيصمد (11 مقابل 9، يعني عملة معدنية). كل اللي غيره — حتى **دمج التلاتة مع بعض** — بيتغلب في الغالبية العظمى.

يعني إيه عملياً؟

- محلل بيختار `AutoARIMA` صح كل مرة ← **مش هيخسر حاجة** لو فضل كلاسيكي.
- محلل **مش** بيعرف يختار، أو معندوش وقت يظبط كل سلسلة ← **أحسن له بكتير** يستخدم الـ foundation model.

من الآخر: الموديل مش بيكسب على أحسن طريقة، لكنه بيكسب على **متوسط** الطرق — وقيمته إنك **مش محتاج تعرف** أنهي طريقة تختار.

Figure 2: Accuracy of TimesFM-3 relative to the best classical method



شكل 2 — دقة `TimesFM-3` بالنسبة لأحسن طريقة كلاسيكية في كل خلية. العمود اللي **تحت** الصفر معناه `TimesFM-3` أدق، واللي **فوق** الصفر معناه الكلاسيكي أدق. شوف `D8` (الطلب المتقطع) نازل تحت بوضوح في الأطوال الأربعة، و `D4` (الموسمية) طالع فوق في الأربعة — دول النظامين الحاسمين اللي هنتكلم عنهم في 8.5. والباقي قريب من الصفر، يعني تعادل.

8.3 ⭐ Specification Is Not Estimability — التخصيص الصحيح مش زي القابلية للتقدير

دي **أهم نتيجة في الورقة كلها**. وعشان تبقى واضحة، خلينا نبني الفكرة من الأول بالراحة قبل ما نشوف أي رقم.

الخطوة 1: يعني إيه "الموديل مخصّص صح"؟

التخصيص (Specification) = **الشكل الرياضي** اللي إنت افترضته.

في المحاكاة بتاعتنا، عملية `D5` بتولّد البيانات من معادلة `ETS(A,A,A)` — يعني مستوى + اتجاه + موسمية. وطريقة `AutoETS` هي بالظبط اللي بتفترض الشكل ده.

يعني إيه؟ يعني **الشكل مظبوط 100%**. مش تقريب ولا اجتهاد — دي **الحقيقة نفسها**. الموديل عارف إن فيه مستوى، وعارف إن فيه اتجاه، وعارف إن فيه موسمية بفترة 12. كل ده **صح**.

فالمفروض يكسب، صح؟ ده اللي إحنا كمان توقّعناه وسجّلناه كفرضية `H1` .

الخطوة 2: بس الشكل الصح مش كفاية — لازم **تعرف الأرقام**

هنا المفتاح اللي بيغيّر كل حاجة.

معرفة الشكل حاجة، ومعرفة **الأرقام اللي جوه الشكل** حاجة تانية خالص. والأرقام دي إنت **لازم تقدّرها من البيانات اللي قدامك**.

فخلينا نعدّ: `ETS(A,A,A)` بموسمية 12 — كام رقم لازم يتقدّر؟

اللي لازم يتقدّر	العدد
معاملات التنعيم: α، β، γ	3
المستوى الابتدائي ℓ_0	1
الاتجاه الابتدائي b_0	1
الحالات الموسمية الابتدائية (11 حرّة من 12)	11
الإجمالي	**16 رقم**

دلوقتي السؤال الحاسم: عندك كام نقطة بيانات عشان تقدّر الـ 16 رقم دول؟

- عند `n = 200` ← عندك **12.5 نقطة لكل رقم**. مرتاح جداً.
- عند `n = 24` ← عندك **1.5 نقطة لكل رقم فقط!**

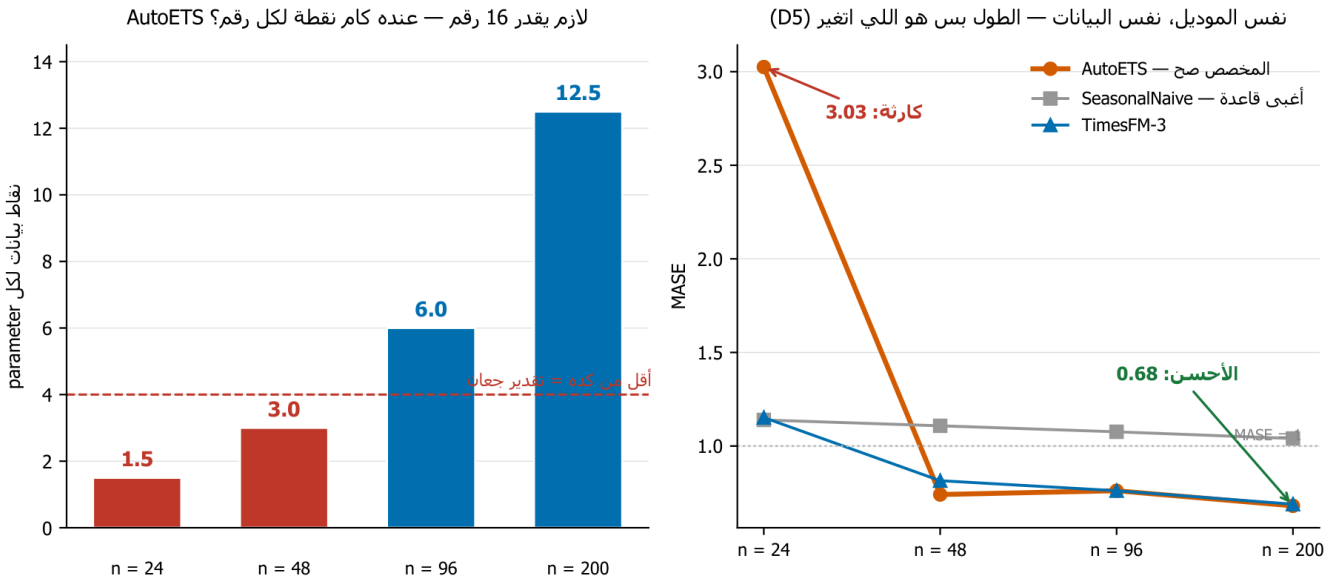
تخيّل الموقف: إنت بتحاول تحدّد 16 مجهول من 24 معادلة. ده رياضياً ممكن بالعافية، بس البيانات فيها ضوضاء (noise)، فاللي بيحصل إن الموديل **بيحفظ الضوضاء** بدل ما يتعلّم النمط — وده اسمه **overfitting**. وبيطلع تنبؤ كارثي.

تشبيه: عندك وصفة أكل مضبوطة 100% من شيف عالمي — الوصفة بتقول "حط 16 نوع بهارات بنسب معيّنة". بس مطبخك فيه **مكوّنين بس**. الوصفة **صح**، إنما إنت **مش قادر تنفّذها**. والراجل اللي بيطبخ بالفطرة من غير وصفة أصلاً هيطلع أكل أحسن منك.

الخطوة 3: الفرق بين الاتنين في جملة

	التخصيص (Specification)	القابلية للتقدير (Estimability)
السؤال	الشكل الرياضي صح؟	عندي بيانات تكفي أحدّد أرقامه؟
بيعتمد على	العملية اللي ولّدت البيانات	**طول السلسلة** اللي قدامي
`D5` على `AutoETS`	**صح دايماً** ✓	صح عند `n≥48` ، **فاشل** عند `n=24`

الخطوة 4: الدليل — نفس الموديل، من الأسوأ للأحسن



شكل 6 — ليه الموديل المخصّص صح ممكن يقع. الشمال: `AutoETS` لازم يقدّر 16 رقم؛ عند `n = 24` عنده **1.5 نقطة لكل رقم** (الأحمر) وعند `n = 200` عنده **12.5** (الأزرق). اليمين: النتيجة — على `D5` ، وهي **العملية اللي `AutoETS` مخصّص ليها صح**، يبدأ **كارثة عند 3.03** (أسوأ واحد في الستة!) وينتهي **الأحسن عند 0.68**.

بصّ على الخط البرتقالي كويس: الموديل **متغيّرش**. البيانات جاية من **نفس العملية** بالظبط. اللي اتغيّر هو **طول السلسلة بس**. ومع ذلك النتيجة اتقلبت من الأسوأ للأحسن.

والأرقام بالتفصيل عند `n = 24` :

الطريقة	MASE على `D5` عند n=24	إيه اللي بتقدّره؟
`AutoETS`	**3.025** ← المخصّص صح!	**16 رقم** من 24 نقطة
`SeasonalNaive`	**1.139** ← أغبى قاعدة	**صفر** — مفيش حاجة تتقدّر
`TimesFM3`	1.152	**صفر** — متدرّب قبل كده

العمود التالت هو التفسير كله. `SeasonalNaive` بتقول "النهارده زي نفس الشهر السنة اللي فاتت" — **مفيش أي رقم بتقدّره**، فمفيش حاجة تغلط فيها. و `TimesFM-3` كمان **مبيقدّرش حاجة** من سلسلتك — هو جاي ومعاه معرفة مسبقة من تريليون نقطة. الاتنين **محصّنين** ضد مشكلة قلة البيانات، و `AutoETS` **مش محصّن**.

يعني `AutoETS` خسر بفارق **2.7 مرة** قدام قاعدة من سطر واحد — **على بيانات عملية ETS هي اللي ولّدتها بنفسها!**

الخطوة 5: نفس الحكاية في `D4`

مش حالة شاذة. في `D4` (الموسمية الـ SARIMA) نفس الحاجة:

- عند `AutoARIMA` = **2.284** : `n = 24` ، و**1.029** = `SeasonalNaive` . الكلاسيكي الأوتوماتيكي المخصّص صح **خسر** قدام القاعدة الغبية.
- عند `AutoARIMA` = **0.886** : `n = 200` ، بقى الأحسن بفارق كبير.

نفس النمط: الشكل صح من الأول لآخر، بس **القدرة على تقديره** هي اللي بتتغيّر مع الطول.

الخطوة 6: وليه ده مهم عملياً؟

الفرضية `H1` اللي سجّلناها كانت بتقول: "الكلاسيكي هيكسب على أرضه، **والفرق هيبقى أكبر في السلاسل القصيرة**". طلعت **مفنّدة** — والفرق في السلاسل القصيرة راح **في الاتجاه المعاكس**.

والدرس اللي تاخده معاك:

> السؤال مش **"الموديل بتاعي صح؟"** — السؤال **"عندي بيانات تكفي أقدّره؟"**

وده بيقولك **فين** الـ foundation models بتنفع بالظبط: **في المكان اللي التقدير الكلاسيكي فيه جعان بيانات**. مش لأنها "أذكى"، لكن لأنها **مش محتاجة تقدّر حاجة أصلاً** من سلسلتك — هي جايّة ومعاها المعرفة جاهزة.

وبالعكس: لما يبقى عندك بيانات كتير ومرتاحة، الكلاسيكي بيرجع يكسب — زي ما شفنا عند `n = 200` .

من الآخر: الموديل الصح + بيانات قليلة = نتيجة وحشة. والقاعدة الغبية + صفر تقدير = نتيجة معقولة. الـ foundation model بيلعب في الخانة التانية دي — بيجيب نتيجة معقولة من غير ما يقدّر أي حاجة.

8.4 Robustness, Not Peak Accuracy — الموديل بيشتري الثبات مش الدقة القصوى

إزاي نقيس الثبات ده بشكل عادل؟

⚠️ **خد بالك: أسوأ MASE مطلق هو المقياس الغلط هنا.** ليه؟ لأن `D3` صعبة على كل الطرق، فأسوأ الحالات المطلقة بتبقى محكومة **بالعملية دي** مش بهشاشة أي طريقة. الدليل: أسوأ حالة مطلقة عند `Theta` هي 4.483 — **أحسن** من بتاعة `TimesFM-3` (4.656) ! ولو مشينا بالمقياس ده يبقى `Theta` أثبت طريقة، وده كلام مايتقالش لما تبصّ على أداءه في `D4` (4.155 مقابل 0.886).

المقياس الصح: نسبة كل طريقة لـ **أحسن طريقة في نفس الخلية**. ده بيعزل الهشاشة لوحدها:

الطريقة	النسبة الوسيطة	المتوسط	أسوأ نسبة	أسوأ خلية
`TimesFM3`	**1.008**	**1.054**	**1.31**	D4, n=200
`AutoARIMA`	1.032	1.137	2.22	D4, n=24
`Combination`	1.108	1.273	3.00	D4, n=24
`AutoETS`	1.194	1.366	3.50	D4, n=24
`SeasonalNaive`	1.200	1.470	5.65	D7, n=48
`Theta`	1.251	1.614	4.69	D4, n=200

اقرا الجدول ده كويس:

- `TimesFM-3` **عمره ما بقى أسوأ من 1.31 مرة** من أحسن طريقة في أي خلية في التصميم كله.
- وفي الخلية الوسيطة هو **في حدود 0.8% من الأحسن**.
- بينما **كل** طريقة كلاسيكية بتبقى في مكان ما **2.2 مرة أسوأ على الأقل**، وبعضها 4.7 و 5.6 مرة.

تشبيه: لاعب بيجيب 8/10 في كل ماتش، مقابل لاعب بيجيب 10/10 مرة و 3/10 مرة. اللي عايز يكسب البطولة كلها بياخد الأول.

من الآخر: `TimesFM-3` نادراً ما بيبقى الأحسن، وأبداً مش بيبقى كارثة. الكلاسيكي الأوتوماتيكي أحياناً ممتاز وأحياناً **بيقع**.

8.5 Two Decisive Regimes — النظامين اللي فيهم قرار واضح

أ) موسمية قوية + بيانات تكفي ← استخدم `AutoARIMA`

عند `n = 48, 96, 200` في `AutoARIMA` ، `D4` أحسن بـ **21.8% و 21.1% و 23.7%** على الترتيب، وكلها بـ p مصحّح < 0.0001.

⚠️ **وخد بالك من قراءة النسبة — دي غلطة شائعة.** النسبة 1.28 معناها إن `TimesFM-3` **أسوأ بـ 28%**، لكنها **مش** معناها إن `AutoARIMA` **أحسن بـ 28%**. دول رقمين مختلفين:

$$\underbrace{1.28 - 1 = 28\%}_{\text{TimesFM-3 worse by}} \quad \neq \quad \underbrace{1 - \frac{1}{1.28} = 22\%}_{\text{AutoARIMA better by}}$$

عشان كده بنقول `AutoARIMA` **أحسن بـ 21–24%**. والقاعدة العامة: لو `أ` أسوأ من `ب` بنسبة r، يبقى `ب` أحسن من `أ` بنسبة $1 - \frac{1}{1+r}$ — وهي **دايماً أصغر** من r.

والتوصية بتنقلب عند `n = 24`! بدورتين موسميتين بس: - أحسن طريقة كلاسيكية مش `AutoARIMA` أصلاً — دي (1.029) `SeasonalNaive` - و `TimesFM-3` **بيكسب على `AutoARIMA` بـ 42%** (1.321 مقابل 2.284)

يعني نصيحة "استخدم ARIMA للموسمية" صح **بس بعد** ما يبقى الموديل الموسمي قابل للتقدير. تحت الحد ده هي نصيحة **وحشة**.

ب) الطلب المتقطع (`D8`) ← الحكاية أعقد، والمقاييس بتتخانق

على `TimesFM-3` ، `MASE` بيقلّل الخطأ بـ **21.7% و 26.5% و 26.0% و 26.2%** على الأطوال الأربعة، وكلها p < 0.0001. وعلى الـ `pinball loss` كمان هو الأحسن في كل طول.

بس على `RMSSE` هو بيخسر في كل طول. وعلى `sMAPE` هو أوحش واحد في الستة بفارق كبير.

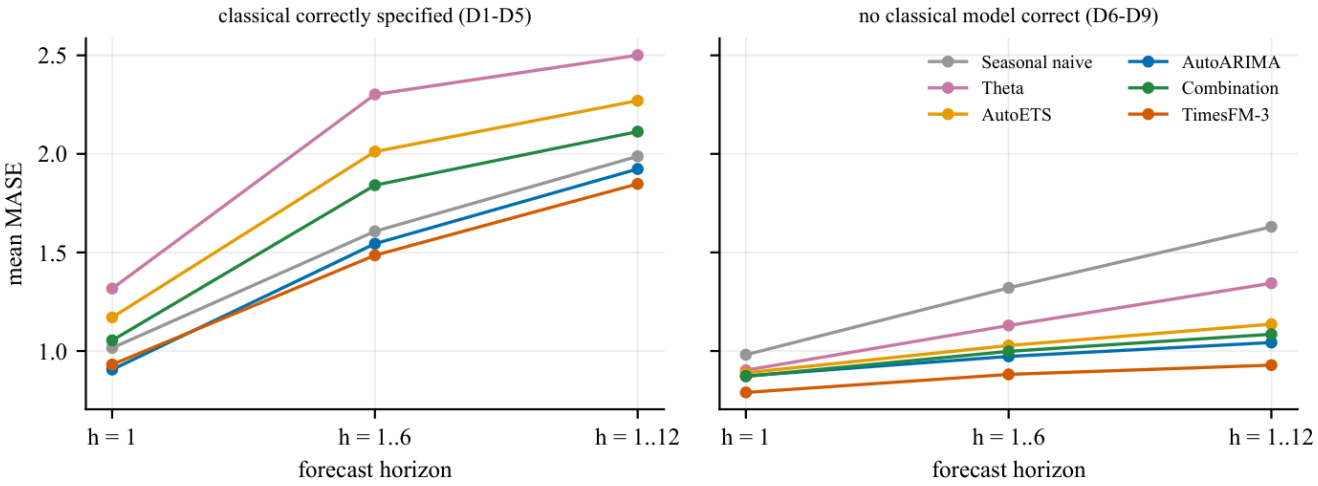
ليه؟ فيه ميكانيزم حقيقي مش صدفة:

70% من قيم `D8` أصفار. والقاعدة الرياضية بتقول: - مقاييس **الخطأ المطلق** (زي `MASE`) بتتصغّر عند الـ **median** بتاع التوزيع — وده هنا صفر أو قريب منه. - مقاييس **الخطأ التربيعي** (زي `RMSSE`) بتتصغّر عند الـ **mean** — وده أكبر من صفر بالتأكيد.

فالطريقة اللي بتتنبأ قريب من الـ median بتكسب MASE وتخسر RMSSE، والعكس صحيح. ودي بالظبط المصيدة اللي `Kolassa (2016)` وثّقها للطلب المتقطع في التجزئة.

من الآخر: في الطلب المتقطع، **مفيش رقم واحد بيحسم الموضوع** — اللي يحسمه هو دالة الخسارة اللي قرارك التجاري بيواجهها فعلاً.

Figure 4: Accuracy by horizon, split by whether a classical model is correct



شكل 4 — الدقة حسب طول الأفق، مقسومة على النظامين. الشمال: العمليات اللي فيها موديل كلاسيكي صح (`D1` – `D5`). اليمين: اللي مفيهاش (`D6` – `D9`). كل الطرق بتسوء كل ما الأفق يطول — وده طبيعي.

⚠️ **بس خد بالك من قراءة غلط هنا:** `TimesFM-3` مبيّن الأقل في **اللوحتين**، وده مش بيناقض إن الكلاسيكي كسب خلايا أكتر — لأن **المتوسط** بيخلط الخلايا اللي الكلاسيكي كسبها بالخلايا اللي **وقع** فيها، وخلية واحدة كارثية (زي `Theta` بـ 4.16 على `D4`) بتحرّك المتوسط أكتر بكتير من كذا مكسب صغير.

8.6 ⭐ Adding the Right Tool: Croston — ضفنا الأدوات المتخصصة فعلاً

المراجعة قالت لنا حاجة وجيهة: "مجموعة المقارنة بتاعتكم **مفيهاش ولا طريقة متصممة للطلب المتقطع**!" وده اعتراض عادل — `Croston` وعيلته موجودين بالظبط للحالة دي.

فضفناهم: `TSB` , `IMAPA` , `ADIDA` , `CrostonSBA` , `CrostonOptimized` , `CrostonClassic` — بنفس الـ seeds ونفس الـ 200 تكرار.

النتيجة (وهي قوية في الاتجاهين):

على `TimesFM-3` : `MASE` **كسب على الستة كلهم في الأطوال الأربعة — 24 من 24 مقارنة معنوية**، بنسب وسيطة بين **0.70 و 0.82**. أحسن منافس متخصص كان `TSB` عند n=96 بـ 0.916، مقابل **0.683** لـ `TimesFM-3` .

على `ADIDA` : `RMSSE` و `CrostonSBA` سجلوا الأقل في **كل** طول (0.694–0.769) مقابل 0.774–0.806 لـ `TimesFM-3` .

والنتيجة دي بتأكد الميكانيزم بشكل أنضف من الأول: طرق `Croston` **متصممة أصلاً** لتقدير **معدل الطلب المتوسط** (mean demand rate) — وده بالظبط اللي الخطأ التربيعي بيكافئه! فالطرفين متحسّنين لدوال مختلفة، وكل واحد بيكسب تحت المقياس اللي بيوافق هدفه. مفيش حد منهم غلط.

فالاعتراض اتقفل: الموديل مش بس بيكسب على الطرق العامة — هو بيكسب على **الأدوات المتخصصة** كمان، وبفارق **أوسع**. لكن على المقياس التربيعي بيخسر لهم.

من الآخر: لو تكلفتك في المخزن **خطية** بالقطعة ← `TimesFM-3` . لو **تربيعية** ← `Croston` . وقول ده بصراحة أحسن من إنك تختار المقياس اللي يعجبك.

8.7 Interval Reliability — هل الفترات بتاعتهم صادقة؟

لما موديل يقولك "أنا واثق 80%"، هل فعلاً الحقيقة بتقع جوه فترته 80% من المرات؟

النظرة المتوسطة (على العمليات التسعة): تغطية `TimesFM-3` للفترة 80% بتفضل بين **0.802 و 0.834** عبر كل الأطوال — أقرب للمطلوب من أي طريقة تانية. بينما `SeasonalNaive` بتنجرف من 0.777 عند n=24 لـ 0.869 عند n=96.

⚠️ **بس النظرة دي مجاملة للكل، وقلنا الرقم الأمين جنبها:** لما تاخد **متوسط** التغطية عبر العمليات، الزيادة في عملية **بتلغي** النقص في عملية تانية. المقياس الصح هو **متوسط الانحرافات المطلقة لكل خلية**، وهو **حوالي 3 أضعاف**:

الطريقة	انحراف "المتوسط أولاً" (مجامل)	الانحراف الأمين لكل خلية
`TimesFM3`	0.015	**0.047**
`AutoARIMA`	0.021	0.066
`Combination`	0.032	0.089
`AutoETS`	0.021	0.101
`Theta`	0.040	0.127
`SeasonalNaive`	0.038	0.144

الترتيب متغيّرش و `TimesFM-3` فضل الأحسن معايرةً — **بس المستوى مش قريب من المطلوب**: تغطيته لكل خلية بتتراوح من **0.583 لـ 0.921**! يعني **مفيش ولا طريقة** في الدراسة دي بتدي فترات 80% موثوقة عبر العمليات التسعة.

الحالة اللي بتوضح كل حاجة: الكسر الهيكلي (D6)

	التغطية (المطلوب 0.80)	عرض الفترة (مقيس)
كل الطرق الكلاسيكية	**0.966 – 0.999**	**4.81 – 10.36**
`TimesFM3`	0.828 – 0.889	**3.00 – 5.07**

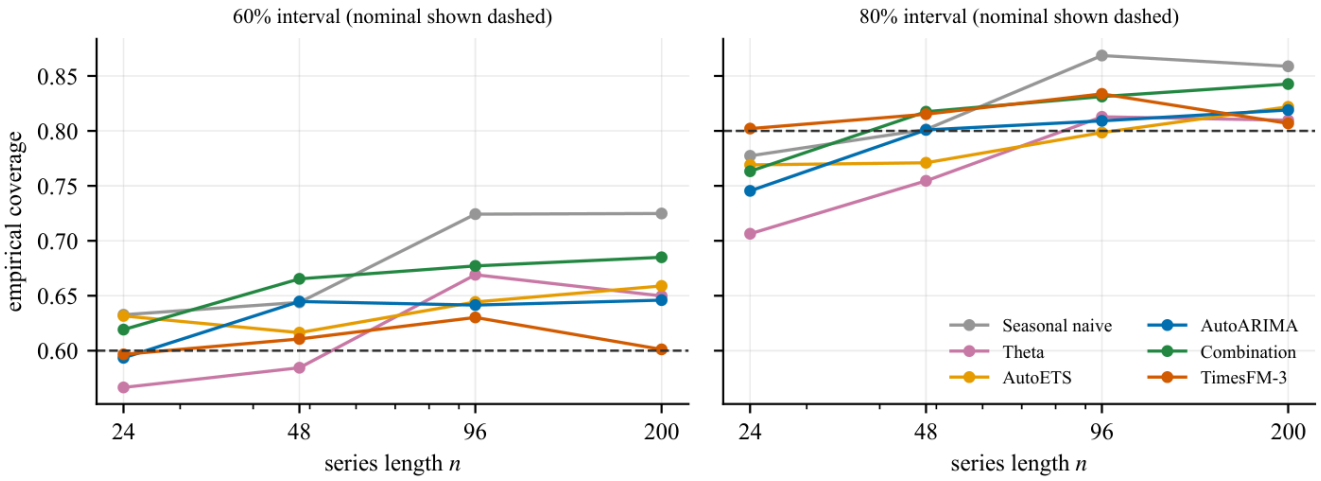
بصّ للكلاسيكي: تغطية 99.9%؟! ده مش "موثوقية" — ده **عرَض مرضي**. الكسر الهيكلي بيضخّم تقدير التباين، فالفترات بتبقى **ضخمة**. هي بتحتوي الحقيقة دايماً لأنها بتحتوي **كل حاجة**.

`TimesFM-3` بيغطي 82.8–88.9% (فوق المطلوب برضه) بفترات **نص العرض تقريباً**.

تشبيه: لو قلتلك "بكرة الجو هيبقى بين -50 و +80 درجة" — كلامي صح 100%، وبلا فايدة 100%.

من الآخر: التغطية لوحدها بتكدب. لازم تبص على **العرض** معاها.

Figure 3: Interval reliability, averaged over all nine processes



شكل 3 — التغطية الفعلية للفترات مقابل المطلوب (الخط المتقطّع)، بمتوسط على العمليات التسعة. لاحظ إن `TimesFM-3` ماشي على الخط المطلوب تقريباً في كل الأطوال، بينما الكلاسيكي **بيقصّر** عند `n = 24` و**بيزوّد** من `n = 96`. ⚠️ بس افتكر التحذير فوق: ده **المتوسط**، والانحراف الحقيقي لكل خلية أكبر 3 مرات.

8.8 Was the Combination a Straw Man? — هل الـ Combination كان خصم ضعيف؟

`TimesFM-3` كسب على الـ Combination في 27 من 30 مقارنة معنوية. وده بيفتح اعتراض وجيه: الـ Combination بتاعكم فيه `Theta` اللي بيقع بشكل كارثي في خلايا معينة، و**المتوسط مش محصّن** ضد عضو واحد بايظ.

فأعدنا المقارنة ضد **combination بالوسيط (median)** بدل المتوسط، وده محصّن ضد عضو فاشل.

النتيجة: الاعتراض مش صامد. الـ median combination طلع **أسوأ شوية** من الـ MASE (mean = 1.712 مقابل 1.656، و1.439 = `TimesFM-3`)، و `TimesFM-3` لسه كاسب 25 من 29 مقارنة معنوية.

يعني تفوّق الموديل على الدمج **مش ناتج** عن طريقة تكوين الدمج. (والتحليل ده **post-hoc/بعدي** وقلنا كده صراحةً.)

8.9 Effect Sizes and a Sensitivity Check — أحجام الأثر وفحص الحساسية

تقديرات `Hodges–Lehmann` (الفرق الوسيط في MASE، السالب لصالح `TimesFM-3`):

الخصم	الفرق الوسيط
`AutoARIMA`	**-0.010**
`Combination`	-0.091
`SeasonalNaive`	-0.133
`AutoETS`	-0.183
`Theta`	-0.227

اقرا الرقم الأول كويس: قدام `AutoARIMA` الفرق النموذجي هو **واحد من مية** من وحدة MASE. ده فرق **بيتقاس** أحياناً، بس **مش فرق محلل هيلاحظه** في شغله.

فحص الحساسية: الـ t-test اتفق مع Wilcoxon في **163 من 180** مقارنة (**90.6%**). والـ 17 اللي اختلفوا فيهم مش عشوائيين — فيهم نمط مهم:

في `D8` قدام `SeasonalNaive` ، الـ **t-test** بيقول `TimesFM-3` أحسن بشكل معنوي في كل طول (p < 0.0001)، بينما **Wilcoxon** بيقول **مفيش فرق معنوي** (p من 0.11 لـ 0.92)!

ليه؟ لأن الاتنين بيجاوبوا على سؤالين مختلفين: - **المتوسط** كبير لأن **أقلية** من السلاسل بتنتج أخطاء ضخمة جداً لـ `SeasonalNaive` . - **الوسيط** قريب من صفر لأن على **معظم** السلاسل الفردية الاتنين متقاربين.

وده **سبب تالت مستقل** يخلينا نعامل نتيجة `D8` كنتيجة **مشروطة** مش حاسمة.

---

9. The Real-Data Tier: M4 Monthly — البيانات الحقيقية

المحاكاة بتشتري النقاء بثمن الواقعية. القسم ده بيرجّع جزء من الثمن ده.

9.1 What Is M4 Anyway? — إيه هي بيانات M4 أصلاً؟

الـ **M4** هي **مسابقة تنبؤ عالمية** نظّمها Spyros Makridakis سنة 2018. حطّوا **100,000 سلسلة زمنية حقيقية** من الاقتصاد والمالية والصناعة والديموغرافيا، وخلّوا الناس كلها تتنافس عليها. دي مش بيانات لعبة — دي المرجع المعياري اللي أي ورقة تنبؤ جادّة بتتقاس عليه.

السلاسل مقسومة بالتردد: سنوية، ربع سنوية، **شهرية**، أسبوعية، يومية، بالساعة. إحنا أخدنا **الشهرية** — وهي أكبر مجموعة فيهم: **48,000 سلسلة**.

ليه الشهرية بالذات؟ لأنها أقرب حاجة للواقع اللي المحلل بيشتغل عليه (مبيعات شهرية، طلبات شهرية، مؤشرات اقتصادية شهرية)، وفيها **موسمية** حقيقية بفترة 12 — وده بيخلّيها الاختبار الطبيعي للنتايج اللي طلعت من `D4` و `D5` في المحاكاة.

9.2 How We Picked the Series — إزاي اخترنا السلاسل بالظبط

مخدناش الـ 48,000 كلهم — لأن ده هياخد أيام حساب من غير فايدة إضافية. المواصفات بالتفصيل:

الخطوة	القيمة	ليه
المجموعة الأصلية	48,000 سلسلة شهرية	كل الشهرية في M4
شرط الطول	السلسلة لازم تكون ≥ **138 شهر**	عشان تكفي لـ context محترم + 18 شهر مختبّيين
المؤهّلة بعد الشرط	**32,581 سلسلة**	اللي عدّت الشرط
العيّنة	**1,000 سلسلة** عشوائي	بـ seed ثابتة `20260908`

ليه seed ثابتة؟ عشان أي حد يشغّل الكود يجيب **نفس الألف سلسلة بالظبط** — مش عيّنة تانية. لو مكانش فيه seed، كل واحد هيجيب نتيجة مختلفة شوية والمقارنة تبقى مالهاش معنى.

طول السلاسل اللي طلعت: الأقصر **138** شهر، الوسيط **299** شهر (يعني حوالي 25 سنة!)، والأطول **1,808** شهر. يعني دي سلاسل **طويلة ومحترمة** مش أمثلة صغيرة.

9.3 ⭐ What Is an "Origin" and a "Window"? — يعني إيه "origin" و"window"؟

دي أهم نقطة محتاجة توضيح، وهي بسيطة جداً لما تتشرح صح.

الـ Origin (نقطة الأصل) = اللحظة اللي بتقف فيها عن المشاهدة وتبدأ تتنبأ.

تخيل إنك واقف في شهر معيّن. كل اللي **قبله** إنت شايفه (ده الـ **context**)، وكل اللي **بعده** مخبّي عنك (ده الـ **truth** اللي هتتقاس عليه). النقطة اللي إنت واقف عندها دي اسمها الـ **origin**.

الـ Window (النافذة) = زوج واحد: (البيانات اللي شفتها، الحقيقة المخبّية).

مثال حقيقي بالأرقام من بياناتنا

خد سلسلة حقيقية من العيّنة، اسمها `M10046` ، طولها **346 شهر**. إحنا بناخد منها **3 نوافذ**:

النافذة	الموديل بيشوف	بيتنبأ بـ	العدد
origin 1	شهور **1** → **328**	شهور **329** → **346**	18 شهر
origin 2	شهور **1** → **316**	شهور **317** → **334**	18 شهر
origin 3	شهور **1** → **304**	شهور **305** → **322**	18 شهر

شوف النمط: كل مرة بنرجع **12 شهر لورا** (سنة كاملة)، وبنطلب منه يتنبأ بـ **18 شهر** من النقطة دي.

ليه بنرجع 12 شهر مش شهر واحد؟ عشان لو رجعنا شهر واحد، النوافذ هتبقى **متداخلة جداً** وكإننا بنقيس نفس الحاجة 3 مرات. الرجوع سنة كاملة بيدّي **مواقف مختلفة فعلاً** في السلسلة.

ليه 3 نوافذ مش واحدة؟ عشان لو قيّمنا على نقطة واحدة بس، ممكن نكون وقعنا على شهر **استثنائي** (أزمة، عيد، حاجة غريبة) والنتيجة كلها تبقى حظ. تلات نقاط بتقلّل الحظ ده.

وليه 18 شهر بالذات؟ لأن ده **الأفق الرسمي** اللي مسابقة M4 نفسها حدّدته للسلاسل الشهرية. يعني نتايجنا تبقى **قابلة للمقارنة** مع أي نتيجة منشورة على M4.

وليه 2,932 نافذة مش 3,000؟

`سلسلة × 3 نوافذ = 3,000 1,000` — بس الرقم الحقيقي **2,932**. الفرق 68 نافذة. ليه؟

السبب في جملة: مش كل سلسلة طولها يسمح بـ 3 نوافذ. السلاسل القصيرة بتدّي أقل، وإحنا **مرميناهاش** — أخدنا منها اللي طولها يسمح بيه.

الشرط بالظبط، ومن فين جه

عشان النافذة رقم k تعدّي (يعني $k = 0$ للأولى، 1 للتانية، 2 للتالتة)، لازم اللي يفضل للموديل يشوفه ما يقلّش عن **120 شهر**. والحساب:

- إحنا بنرجع لورا $12k$ شهر عشان نوصل للـ origin بتاع النافذة دي.
- وبنشيل **18 شهر** (الأفق) عشان نخبّيهم.
- فاللي بيفضل للموديل = $n - 12k - 18$.

فالشرط:

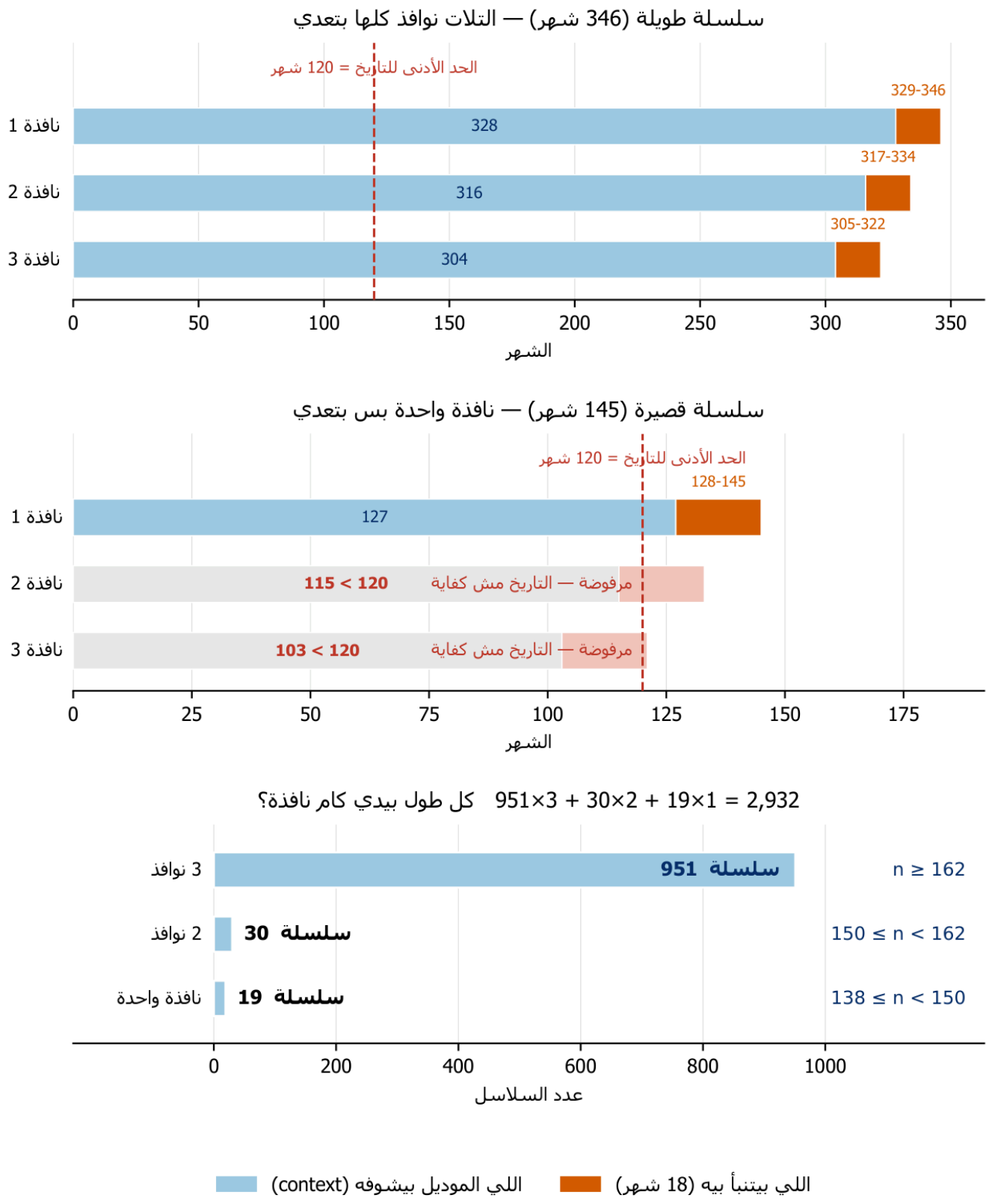
$$n - 12k - 18 \;\geq\; 120 \qquad \Longleftrightarrow \qquad n \;\geq\; 138 + 12k$$

نطبّقه على النوافذ التلاتة:

النافذة	k	الشرط	يعني السلسلة لازم تكون
الأولى	0	$n \geq 138 + 0$	**≤ 138 شهر**
التانية	1	$n \geq 138 + 12$	**≤ 150 شهر**
التالتة	2	$n \geq 138 + 24$	**≤ 162 شهر**

ودي بالظبط الأرقام اللي في الشكل تحت. ولاحظ إن شرط النافذة الأولى (138) هو **نفسه** شرط اختيار العيّنة من الأول — عشان كده مفيش ولا سلسلة طلعت بصفر نوافذ.

الشكل: إزاي السلسلة بتتقسّم لنوافذ



شكل 5 — إزاي السلسلة الواحدة بتتحوّل لنوافذ، وليه القصيرة بتدّي أقل. اللون الأزرق = اللي الموديل بيشوفه، والبرتقالي = الـ 18 شهر اللي بيتنبأ بيهم. الخط الأحمر المتقطّع = الحد الأدنى للتاريخ (120 شهر).

- **فوق:** سلسلة طولها **346 شهر**. كل مرة بنرجع 12 شهر لورا، ولسه فاضل للموديل 328 ثم 316 ثم 304 شهر — كلهم فوق الـ 120 بكتير. فالتلات نوافذ **بتعدّي**.
- **في النص:** سلسلة طولها **145 شهر** بس. النافذة الأولى تمام (بيفضل 127 شهر). لكن التانية هيفضل فيها **115 شهر** — أقل من 120، فـ **مرفوضة**. والتالتة 103، مرفوضة كمان. فالسلسلة دي بتدّي **نافذة واحدة** وخلاص.

- **تحت:** التوزيع النهائي على الألف سلسلة.

الحساب النهائي

عدد النوافذ	شرط الطول	عدد السلاسل	إجمالي النوافذ
3 نوافذ	$n \geq 162$	**951**	$951 \times 3 = 2853$
نافذتين	$150 \leq n < 162$	30	$30 \times 2 = 60$
نافذة واحدة	$138 \leq n < 150$	19	$19 \times 1 = 19$
الإجمالي		**1,000**	**2,932**

$$2853 + 60 + 19 = 2932$$

⭐ وليه ده مهم أصلاً؟

ممكن تقول: "طب ما كنت تشيل الـ 49 سلسلة القصيرة وتخلّيها 951 × 3 = 2,853 وخلاص، الرقم يبقى مظبوط!" — وده بالظبط اللي **معملناهوش**، وليه سببين:

1. **رمي البيانات القصيرة بيحيّز النتيجة.** السلاسل القصيرة هي أصعب حالة على الطرق الكلاسيكية (فاكر درس **الـ estimability** في القسم 8.3؟). لو شلناها، نبقى شلنا بالظبط الحالات اللي `TimesFM-3` بيتفوّق فيها — يعني **حيّزنا النتيجة ضد نفسنا** من غير ما ناخد بالنا.

2. **الرقم "الوحش" أصدق من الرقم المظبوط.** لما تلاقي في ورقة رقم زي 2,932 بدل 3,000 المستديرة، ده مؤشر إن الباحث **مزوّقش** — وصف اللي حصل فعلاً. الأرقام المستديرة أوي في الأوراق أحياناً بتبقى علامة إن حد رمى بيانات عشان الجدول يبقى شكله حلو.

من الآخر: الرقم 2,932 مش غلط ولا نقص — ده **نتيجة قاعدة واضحة** (`n ≥ 138 + 12k`) اتطبّقت على أطوال حقيقية مختلفة، وكل سلسلة أخدت اللي طولها يسمح بيه بالظبط.

9.4 ⭐ ⭐ Why Use Data the Model May Have Seen? — ليه نستخدم بيانات الموديل يمكن يكون شافها؟

ده أهم سؤال في القسم كله، وسؤال وجيه جداً. الإجابة من تلات حتت:

الحتة الأولى: إحنا **مش** بنبني الادعاء الأساسي عليها

الادعاء الرئيسي للورقة مبني على **المحاكاة** — لأن هناك التلوث **مستحيل بالبناء**. القسم ده بيجاوب على سؤال **تاني خالص وأضيق**:

> **السلاسل الحقيقية بتشبه أنهي نظام من الأنظمة التسعة اللي حاكيناها؟**

وده سؤال مشروع تماماً حتى لو البيانات ملوّثة، لأنه سؤال عن **شكل البيانات** مش عن **مين أدق**.

الحتة التانية: المحاكاة لوحدها ممكن تكون **منفصلة عن الواقع**

لو الورقة اكتفت بالمحاكاة، أول اعتراض من أي محكّم هيبقى: *"كلامكم حلو على بيانات صناعية — بس الدنيا الحقيقية مش كده"*. فلازم نوريه إن نتايجنا **بتتكرر** على بيانات حقيقية.

وده اللي حصل بالظبط: نفس النمط ظهر في الاتنين. والاتفاق ده **قيمته أكبر من أي طبقة لوحدها**، لأن **ضعف كل طبقة معاكس للتانية**:

	المحاكاة	M4 الحقيقية
التلوث	**مستحيل** بالبناء	**محتمل جداً**
الواقعية	صناعية ومبسّطة	**حقيقية 100%**

يعني لو حاجة طلعت صح في **الاتنين**، تبقى مش صدفة تصميم ولا صدفة تلوث.

الحتة التالتة ⭐ — التعادل نفسه **دليل**

دي النقطة الذكية اللي تستاهل تفتكرها.

لو كان `TimesFM-3` بيغش فعلاً (يعني حافظ سلاسل M4 من التدريب)، كنا هنتوقع إيه؟ كنا هنتوقع إنه **يسحق** الطرق الكلاسيكية سحق — لأنه حرفياً شايف الإجابة.

بس ده مش اللي حصل. هو **اتعادل** مع `p = 0.27` (52.6%، `AutoARIMA`) واتعادل مع الـ (48.5%، `Combination`) `p = 0.27`). يعني حتى على البيانات اللي هو "متّهم" إنه شافها، **مكسبش عليهم**.

فالتعادل ده في حد ذاته **دليل ضد** فرضية إن الموديل بيستفيد استفادة ضخمة من الحفظ. مش دليل قاطع — يمكن حافظ حتة صغيرة — بس هو **أقوى حاجة** تقدر تستنتجها من بيانات عامة.

من الآخر: إحنا بنستخدم M4 عارفين إنها ملوّثة، وقايلين كده بصراحة في الورقة، ومش بانيين عليها الادعاء الرئيسي. بنستخدمها عشان **نعرف شكل الواقع** ونشوف هل المحاكاة بتوصف الدنيا الحقيقية ولا لأ. والنتيجة إن **الاتنين بيقولوا نفس الكلام**.

النتيجة

الطريقة	متوسط MASE	التغطية 80%	`pinball`
`Combination`	**0.9027**	**0.8006**	0.3653
`TimesFM3`	0.9114	0.7655	**0.3627**
`AutoARIMA`	0.9182	0.7851	0.3744
`AutoETS`	0.9465	0.7909	0.3840
`Theta`	0.9672	0.7344	0.3977
`SeasonalNaive`	1.2768	0.7867	0.5135

والاختبارات (Wilcoxon مزدوج عبر الـ 1,000 سلسلة، مصحّح بـ BH):

المقارنة	% سلاسل TimesFM-3 أحسن فيها	p مصحّح	الحكم
ضد `SeasonalNaive`	**80.1%**	< 0.0001	**TimesFM-3 كسب**
ضد `Theta`	56.8%	< 0.0001	**TimesFM-3 كسب**
ضد `AutoETS`	54.1%	0.0043	**TimesFM-3 كسب**
ضد `AutoARIMA`	52.6%	0.27	**تعادل**
ضد `Combination`	48.5%	0.27	**تعادل**

ملاحظتين مهمتين

الأولى — الطبقتين بيتفقوا. النمط ده **نفس** نمط المحاكاة بالظبط: بيكسب على كل حاجة ما عدا `AutoARIMA` اللي بيتعادل معاه. وسلاسل `M4` الشهرية **طويلة وموسمية** — يعني بالظبط النظام "القابل للتقدير" اللي المحاكاة اتنبأت إن الكلاسيكي هيبقى منافس فيه. والاتفاق ده قيمته أكبر من أي طبقة لوحدها، لأن الاتنين **ضعفهم معاكس لبعض**: المحاكاة نضيفة بس صناعية، والحقيقية واقعية بس ملوثة.

التانية — نتيجة الفترات مانتقلتش. على البيانات المولّدة كان `TimesFM-3` الأحسن معايرةً؛ هنا هو **تاني أسوأ واحد**: تغطية **0.765** مقابل مطلوب 0.80، بينما الـ `Combination` وصل **0.801**. يعني على سلاسل حقيقية فيها بنية مش موجودة في عملياتنا التسعة، الـ quantiles المتعلَّمة بتاعته **بتقصّر**. وده بيخفّف من نتيجة المعايرة اللي فوق، ومتسق مع اللي `Adler (2026) وزمايله` بلّغوا عنه.

لكن لاحظ: على الـ `pinball loss` — وهو proper scoring rule ومسجّل مسبقاً — `TimesFM-3` طلع **الأحسن** (0.3627). فالصورة مش "فترات وحشة" على طول، دي "التغطية وحشة والتوزيع ككل كويس".

من الآخر: على بيانات حقيقية طويلة وموسمية، الموديل والكلاسيكي **متعادلين** في الدقة، والكلاسيكي أحسن في صدق الفترات.

---

10. Costs That Don't Show in an Accuracy Table — تكاليف مش بتظهر في جدول الدقة

10.1 ⭐ Speed — والنتيجة بالعكس تماماً من المتوقع

الطبيعي إنك تفتكر إن موديل بـ 330 مليون parameter هو الغالي، والموديل الإحصائي بـ 3 parameters هو الرخيص. **العكس تماماً** على الجهاز ده:

الطريقة	زمن لكل سلسلة	النسبة لـ TimesFM-3
`SeasonalNaive`	3.1 ms	0.13×
`Theta`	9.0 ms	0.39×
`TimesFM3` (GPU, batched)	**23.2 ms**	**1×**
`AutoETS`	102.0 ms	4.4×
`AutoARIMA`	**2382.7 ms**	**102.8×**

`AutoARIMA` أبطأ 103 مرة من `TimesFM-3` لكل سلسلة!

ليه؟ السبب بنيوي مش عرضي: الـ `AutoARIMA` بيعمل **بحث تدريجي** على الـ orders، وبيركّب كل مرشّح بـ maximum likelihood، **لكل سلسلة لوحدها**. بينما `TimesFM-3` بيعمل **مرور أمامي واحد مجمّع (batched)** وبيوزّع تكلفته على الدفعة كلها.

⚠️ بس التحفظ التاني أهم من الرقم الرئيسي

الرقم ده لكل سلسلة على نواة واحدة — والصورة بتتغيّر لما تشغّل الكل بالتوازي. الأرقام الحقيقية من الـ log بتاع التشغيل:

- الطرق الكلاسيكية الأربعة على الـ 7,200 سلسلة كلها بـ 22 نواة: **7.3 دقيقة**
- `TimesFM-3` على نفس الشغل بـ GPU واحد: **3.2 دقيقة**
- النسبة: **2.3×** — مش 103×!

يعني: الـ 103× صح **لكل سلسلة على نواة واحدة**، لكن الكلاسيكي **بيتوازى ببساطة**، والتوازي بيسترد معظم الفرق. الرقمين بيجاوبوا على سؤالين مختلفين، وإني أقول الأول بس يبقى **مبالغة**. (وكمان `TimesFM-3` محتاج GPU أصلاً.)

من الآخر: الموديل الجبار **مش** الخيار الغالي — لا لكل سلسلة ولا في الإجمالي. بس الفرق الإجمالي 2.3× مش 103×.

10.2 ⭐ ⭐ Licensing — وده أهم بند عملي في الورقة كلها

أوزان `TimesFM-3` صادرة تحت رخصة `timesfm-non-commercial-license-v1.0` — يعني استخدام غير تجاري وغير إنتاجي فقط.

الكود نفسه Apache-2.0، وأوزان `TimesFM-2.5` كانت Apache-2.0 كمان. لكن أوزان النسخة 3 **لأ**.

النتيجة العملية مطلقة وسابقة لأي كلام عن الدقة: لو إنت بتشتغل في تطبيق تنبؤ **تجاري**، فـ `TimesFM-3` بشكله المنشور **مش متاح لك أصلاً — عند أي مستوى دقة**. تقليل 26% في الخطأ على الطلب المتقطع **ملهوش أي معنى** لتاجر تجزئة **مش مسموح له قانونياً** ينشر الموديل ده.

الخيارات المتاحة: تستخدم `TimesFM-2.5` تحت Apache-2.0، أو تتفاوض على شروط منفصلة مع Google، أو تستخدم طريقة كلاسيكية. **كل** طريقة كلاسيكية في المقارنة دي **خالية من القيود دي**.

بنقول ده مش كنقد لـ Google — بنقوله لأنه **بيتشال دايماً** من مقارنات الموديلات، وهو **بيقلب الترتيب رأساً على عقب** لفئة كاملة من المستخدمين.

من الآخر: قبل ما تسأل "أدق؟" اسأل "مسموح لي أصلاً؟"

10.3 Interpretability — القابلية للتفسير والتشخيص

`ARIMA` بيقولك الـ orders والمعاملات والأخطاء المعيارية بتاعتها. `ETS` بيقولك معاملات التنعيم والحالات. ودول بيسمحوا بشغل التشخيص العادي: تفحص الـ residuals، تختبر الارتباط الذاتي المتبقي، تفسّر لصاحب القرار **ليه** التنبؤ بيلف هنا، وتكتشف إن الموديل **بطّل يظبط**.

`TimesFM-3` **مبيدّيش حاجة من دول**. لما يتنبأ غلط، مفيش parameter تبصله ولا تشخيص تشغّله — اللي تقدر تعمله إنك **تلاحظ** الأخطاء وتغيّر الطريقة. وفي بيئة منظّمة (regulated) لازم تبرّر التنبؤ مش بس تنتجه — الفرق ده لوحده ممكن يترجّح على المقارنة كلها.

10.4 Prediction Intervals in Theory — الفترات نظرياً

الفترات الكلاسيكية مشتقّة من **بنية الخطأ المفترضة**، فتقدر تناقشها تحليلياً. أما quantiles `TimesFM-3` فهي **مخرجات متعلَّمة** من رأس شبكة، من غير نظرية. أثبتنا فوق إنها في الواقع أحسن معايرةً في المتوسط — **بس تجريبياً، على العمليات دي**، مش بالبناء. وسلوكها على عملية شبيهة بحاجة مش موجودة في corpus التدريب **مجهول بالتعريف**.

---

11. ⭐ When to Use Which — امتى تستخدم إيه (جدول القرار)

ده خلاصة الورقة العملية:

الموقف	استخدم	الدليل
نشر تجاري أو إنتاجي	كلاسيكي (أو TimesFM-2.5)	أوزان v3 رخصتها غير تجارية — الدقة مش القيد أصلاً
طلب متقطع، وتكلفتك خطية	**TimesFM-3**	كسب على الستة Croston في 24/24 مقارنة
طلب متقطع، وتكلفتك تربيعية	`Croston-SBA` / `ADIDA`	بيكسبوا على RMSSE في كل طول
موسمية قوية + ≥ 4 دورات	**AutoARIMA**	أحسن بـ 21–24%، كلها p<0.0001
موسمية بس ≥ دورتين	SeasonalNaive أو TimesFM-3	الكلاسيكي الأوتوماتيكي بيقع: AutoETS = 3.03
آلاف السلاسل، مفيش وقت تظبط كل واحدة	**TimesFM-3**	بيكسب على كل الطرق ما عدا AutoARIMA
سلسلة واحدة وإنت عارف تحدد وتقدّر الموديل	كلاسيكي	AutoARIMA متعادل (9–11)، ومجاني ومفسَّر
كسر هيكلي قريب والفترات مهمة	**TimesFM-3**	تغطية مشابهة بفترات نص العرض (محاكاة بس)
الإنتاجية مهمة وعندك GPU	**TimesFM-3**	أسرع 103× لكل سلسلة، 2.3× إجمالاً
فترات موثوقة على بيانات حقيقية	الـ Combination	على M4: 0.801 مقابل 0.765 لـ TimesFM-3
لازم تفسّر أو تراجَع (audit)	كلاسيكي	TimesFM-3 مبيدّيش parameters ولا تشخيص

---

12. The Hypotheses — الفرضيات: مفيش واحدة نجت سليمة

سجّلنا 5 فرضيات **قبل** أي تنبؤ. والنتيجة:

#	الفرضية	الحكم
H1	الكلاسيكي يكسب على أرضه، والفرق أكبر عند n صغير	❌ **مفنّدة** — الفرق بيروح في الاتجاه المعاكس عند n صغير
H2	TimesFM-3 يكسب حيث لا شكل كلاسيكي صحيح	⚠️ **مدعومة جزئياً** — على MASE و pinball نعم، على RMSSE و sMAPE لأ
H3	في D9 يتعادلوا في المتوسط والكلاسيكي يقصّر في التغطية	⚠️ **جزئياً** — التعادل صح، لكن AutoARIMA بس هو اللي بيقصّر
H4	الـ Combination يكسب على كل فرد ويكون أصعب خصم	❌ **مفنّدة في الشقين** — AutoARIMA أحسن منه، وهو الخصم الأصعب (9–11 مقابل 3–27)
H5	الطرفين يقصّروا في تغطية 80%	❌ **مفنّدة** — على المتوسط TimesFM-3 على المطلوب والكلاسيكي **بيزوّد** من n=96

النقطة المهمة: إن **كل** فرضية فشلت في حتة ما هو **أقوى حجة** للتسجيل المسبق. لو الفرضيات اتكتبت **بعد** النتايج، القسم ده كان هيبلّغ **5 تأكيدات**.

من الآخر: التسجيل المسبق مش بيروقراطية — هو اللي خلانا نقول "غلطنا" 3 مرات بدل ما نقول "توقعنا صح" 5 مرات.

---

13. Limitations — حدود الدراسة (بصراحة)

1. **قوة المحاكاة هي حدودها كمان.** سلاسل مولّدة من 9 عمليات بارامترية **مش** سلاسل حقيقية: هي نضيفة، ومستقرة حيث حددنا، وخالية من خطأ القياس والتأثيرات التقويمية والعروض الترويجية وتغيّرات النظام. اللي اشترته المحاكاة (غياب التلوث + معرفة التخصيص الصحيح) دفعت تمنه **واقعية**.
2. **قيّمنا الموديل zero-shot و univariate بس.** الجديد الأساسي فيه هو الـ **multivariate** بالـ covariates، والقدرة دي **مش متختبرة هنا** لأن عملياتنا كلها أحادية المتغير بالبناء. دراسة تستغل المعلومات عبر السلاسل ممكن توصل لنتايج مختلفة — وده الامتداد الطبيعي.
3. **مفيش تظبيط يدوي لأي طريقة.** ده مقصود وهو الوضع اللي معظم المحللين بيقابلوه، لكن خبير يشوف كل عملية بعينه ويركّب الموديل المناسب بإيده **هيكسب على كل** طريقة أوتوماتيكية هنا.
4. **الأفق قصير** (H=12) والفترة الموسمية ثابتة عند 12. آفاق أطول وترددات تانية ممكن تختلف.
5. **تحليل الطلب المتقطع بعدي (post-hoc)** — أضفنا Croston بعد ما شفنا النتايج، رداً على ثغرة اتحددت في المراجعة، وبنبلّغه منفصل عن المقارنة المسجّلة مسبقاً.
6. **دي دراسة عن موديل واحد.** `Chronos` و `Moirai` مش متقيّمين هنا، فمفيش حاجة هنا تتقري كادعاء عن الـ foundation models عموماً.

---

14. Honesty & Reproducibility — إعادة الإنتاج والأمانة

14.1 Pre-registration and Deviations — التسجيل المسبق والانحرافات

- `PREREGISTRATION.md` **متجمّد قبل** أي تنبؤ، و**متعدّلش أبداً** بعد كده.
- `DEVIATIONS.md` فيه **9 انحرافات**، كل واحد بتاريخه وسببه، **مقسومين مجموعتين**:
- **D-01 لـ D-03:** اتعملوا **قبل** وجود أي نتيجة (تصليح توزيعات البيانات).
- **D-04 لـ D-09:** اتعملوا **بعد** وجود نتايج، وكل واحد بيقول كده في مدخله.

⚠️ **والتقسيمة دي مهمة، مش شكلية:** انحراف اتعمل **قبل** ما تشوف أي نتيجة مختلف تماماً عن انحراف اتعمل **بعدها**. الأول مستحيل يكون متحيّز لنتيجة معيّنة، لأنها لسه مش موجودة أصلاً؛ التاني **ممكن** يكون. عشان كده كل مدخل في المجموعة التانية بيقول صراحةً إنه بعدي، وإيه اللي دفعه، وهل بيأثر على نتيجة مسجّلة مسبقاً ولا لأ.

14.2 Real Bugs Caught — أخطاء حقيقية اتمسكت، وإزاي

⭐ **الغلط الأخطر: اختلال محاذاة صامت.** لما رتّبنا سلاسل `M4` بـ `unique_id = str(i)`، مكتبة `statsforecast` رجّعت الصفوف بترتيب **أبجدي**: "0", "1", "10", "100", "1000"... بينما الكود بيعيد التشكيل على أساس ترتيب **رقمي**. النتيجة: **كل تنبؤ اتقرن بالسلسلة الغلط**. و**مفيش أي رسالة خطأ** — الكود اشتغل عادي!

اكتشفناه إزاي؟ لأن `SeasonalNaive` طلع MASE = **28.9**، وده **مستحيل**: هو أساساً مقام الـ MASE، فلازم يبقى حوالي **1**.

الدرس (وده أهم درس هندسي في المشروع): لازم يبقى معاك دايماً **رقم واحد على الأقل تعرف قيمته الصحيحة مسبقاً**، عشان الغلط الصامت يبقى **مستحيل** مش بس **مستغرب**. والحل: `f"{i:07d}"` — أصفار في الأول عشان الترتيب الأبجدي = الرقمي.

غلط تاني: الـ `Diebold–Mariano` غير الصالح (شرحناه في 7.2) — اتمسح.

غلط تالت: ادعاء إن "كل DOI اتحقق منه عبر Crossref" كان **غير دقيق** — السكربت كان بيراجع قايمة **مكتوبة بالإيد** موازية للـ `refs.bib`، والاتنين افترقوا، و**6 DOIs** في المراجع مراجعوش خالص (كلهم arXiv، وCrossref مبيخدمهاش أصلاً). اتصلّح: السكربت دلوقتي **بيقرا الـ bib نفسه** وبيحلّ الـ DOI عبر `doi.org` اللي بيرد عن Crossref و DataCite. دلوقتي **30 من 30** بيتحلّوا.

14.3 Archiving and Publication — الأرشفة والنشر

- **المستودع العام:** https://github.com/ebrahimkhaled/timesfm-classical-paper (172 ملف، 115 ميجا: الكود كامل + المقاييس لكل سلسلة في الـ 129,600 تقييم + التنبؤات الخام + الأشكال + مصدر المخطوطة).
- **Zenodo:** بالطريقة المعتادة — تفعيل المستودع من موقع Zenodo، وبعدين عمل release على GitHub، والـ webhook بيسك الـ DOI لوحده. وملف `.zenodo.json` موجود في جذر المستودع فـ Zenodo بياخد العنوان والـ ORCID والرخصة منه أوتوماتيك.
- **ليه ننشر التنبؤات الخام؟** مش للكمال — عشان نتيجة الطلب المتقطع **بتنقلب** بين المقاييس، فأي قارئ عايز يعيد التقييم بدالة خسارة مختلفة يقدر **من غير ما يعيد الدراسة كلها**.
- **arXiv:** الحزمة اتبنت واتحققت في **غرفة نظيفة** (extract في مجلد فاضي + compile) — صفر أخطاء، صفر مراجع غير معرّفة. الفئة الأساسية `stat.ME` مع cross-list على `stat.AP` و `cs.LG`.

- **أوزان الموديل مش متعادة نشرها** — رخصتها غير تجارية، والكود بينزّلها من Hugging Face.

---

15. الخلاصة الكبيرة — لو هتفتكر 5 حاجات بس

1. **مفيش طرف كسب.** الموديل كسب 16 خلية والكلاسيكي 20 — بس الـ 20 دول موزعين على **5 طرق مختلفة**، والموديل **واحد من غير تظبيط**.
2. `AutoARIMA` **هو الوحيد اللي بيصمد** (9–11). كل اللي غيره — حتى دمج التلاتة — بيتغلب في الغالبية.
3. **التخصيص الصحيح ≠ القابلية للتقدير.** بدورتين موسميتين، `AutoETS` خسر بـ **2.7 مرة** قدام قاعدة من سطر واحد — **على بيانات هو اللي ولّدها**. الـ foundation models أنفع بالظبط لما التقدير الكلاسيكي جعان بيانات.
4. **الموديل بيشتري الثبات مش الدقة القصوى.** عمره ما بقى أسوأ من **1.31×** من الأحسن، بينما كل كلاسيكي بيبقى ≤ **2.2×** أسوأ في مكان ما.
5. **حاجتين بيقطعوا عرض المقارنة كلها:** الموديل **أسرع 103×** لكل سلسلة من `AutoARIMA` (مش أغلى!)، و**رخصته غير تجارية** — يعني لفئة كاملة من المستخدمين السؤال محسوم **قبل** أي كلام عن دقة.

من الآخر خالص: مش "الجديد أحسن من القديم" ولا العكس — الجديد **أثبت**، والقديم **أدق لما يقدر يتقدّر**، والرخصة ممكن تحسم الموضوع قبل الاتنين.

الملف ده شرح مصاحب للورقة. الأرقام كلها متطابقة مع ملفات النتايج في `results/` ، وكل رقم في الورقة نفسها متربوط بملف CSV بيثبته (18 من 18 فحص بيعدّي).