تقدير القيم المفقودة بالتطبيق على بيانات الوفيات

دراسة مقارنة

إعداد

محمد عبد اللطيف زايد

أستاذ مساعد بقسم الأساليب الكمية

كلية إدارة الأعمال - جامعة الملك فيصل

مدرس بقسم الإحصاء التطبيقي والتأمين

كلية التجارة - جامعة المنصورة

m.a.zayed@mans.edu.eg

المجلة العلمية للدراسات والبحوث المالية والتجارية

كلية التجارة – جامعة دمياط

المجلد الرابع - العدد الأول – الجزء الرابع - يناير ٢٠٢٣

التوثيق المقترح وفقاً لنظام APA:

الكردي، محمد صلاح محمد غريب (٢٠٢٣). تقدير القيم المفقودة بالتطبيق على بيانات الوفيات: دراسة مقارنة. **المجلة العلمية للدراسات والبحوث المالية والتجارية**، كلية التجارة، جامعة دمياط، ٤(١)٤، ٦٤٣-٦٦٦.

رابط المجلة: https://cfdj.journals.ekb.eg/

تقدير القيم المفقودة بالتطبيق على بيانات الوفيات: دراسة مقارنة

محمد عبد اللطيف زايد

الملخص:

يتطلب تقدير الأقساط في تأمينات الأشخاص وكذلك إعداد الخطط السكانية توفر بيانات دقيقة ومكتملة عن كل من الوفيات والسكان في مختلف الأعمار. وفي بعض الأحوال، قد يكون هناك فقد أو قيم مفقودة في البيانات مما يجعل تقدير أو استكمال تلك القيم المفقودة من الأمور الهامة في العلوم الاكتوارية وعلم السكان وموضع اهتمام كثير من الباحثين المختصين. وتهتم هذه الدراسة بالمقارنة بين عدة طرق رياضية وإحصائية معلمية ولامعلمية، مثل الاستكمال الخطي والاستكمال التكعيبي المتدرج لهيرمت Hermite باستخدام كثيرات الحدود (PCHIP) وشرائح التمهيد المقطعة، لاستكمال البيانات المفقودة سواء في أعداد الوفيات أو معدلات الوفاة. وقد تم تطبيق تلك الطرق على بيانات الوفيات لسبع دول أوروبية عن الفترة 2018-2020 مع افتراض عدة حالات للقيم المفقودة. وقد قدمت الطرق المستخدمة فيما عدا الاستكمال بكثيرات الحدود نتائج مقبولة في أغلب الحالات، مع الأخذ في الاعتبار أن بعض الطرق قد تكون هي الأنسب في حالة البيانات المتجانسة وعندما تكون نقاط البيانات غير متباعدة. وكذلك يفضل استخدام الطرق التي تقوم في الأساس على تمهيد البيانات إذا كان الهدف هو الحصول على قيم ممهدة. وقد تمت التوصية بعمل مقرنات بين الطرق التي طبقت في هذا البحث وغيرها، لتقدير القيم المفقودة في بيانات الوفيات عند الأعمار الصغيرة والكبيرة، وخاصة عند حدود البيانات، وكذلك في حالة السلاسل الزمنية المقطعية.

الكلمات المفتاحية: الاستكمال الخطي؛ الاستكمال التكعيبي المتدرج لهيرمت Hermite باستخدام كثيرات الحدود (PCHIP)؛ الشرائح التكعيبية؛ شرائح التمهيد المقطعة

1. مقدمة:

تعد بيانات الوفيات هي الأســـاس في إعداد جداول الحياة والوفاة القومية والاكتوارية، والتي تعتبر من الأسس التي يعتمد عليها في التخطيط السكاني وكذلك تقدير الأقساط في تأمينات الأشخاص بمختلف أنواعها. وفي بعض الأحيان، قد تحتوي البيانات الأصـلية على بعض القيم المفقودة، خاصـة إذا كانت البيانات قد تم تجميعها على أساس آحاد الأعمار، وبالتالي يعتبر تقدير أو استكمال تلك القيم المفقودة من الأمور الهامة في العلوم الاكتوارية وعلم الســكان وموضـــع اهتمام كثير من الباحثين المختصين.

وهناك العديد من الأساليب التي يمكن استخدامها في تقدير واستكمال بيانات الوفيات، خاصة أعداد الوفيات، يمكن تصنيفها إلى فئتين رئيسيتين: الاستكمال بالطرق البيانية، والاستكمال باستخدام الصيغ الرياضية، ولهذه الأخيرة أشكال متعددة بعضـها رياضـي وبعضـها معلمي والآخر لا معلمي. ومن أمثلة ذلك: الصيغ القائمة على القيم المحورية، المتوسطات المتحركة، توفيق المنحنيات والدوال كثيرة الحدود، وبعض الصـيغ الرياضـية الخاصـة التي يعتمد أغلبها على نماذج انحدار لامعلمية مثل الشرائح التكعيبية وشرائح التمهيد الجزائية وغيرها.

وقد تطورت بشــكل كبير طرق اســتكمال وتقدير البيانات المفقودة، وبشــكل خاص تلك التي تعتمد على شـــرائح التمهيد، وقد يرجع الســـبب في ذلك إلى أن الطرق اللامعلمية، ومنها شـــرائح التمهيد، لا تفترض أي قيود حول توزيع البيانات وإمكانية تقدير مصـــفوفة التباين والتغاير وشـــكل التوزيع وغير ذلك من الفرضيات التي تقيد استخدام الطرق المعلمية، كما أنها تعطي تقديرات متسقة مع البيانات الأصـــلية. لذا صـــار اســـتخدام هذه الطرق شـــائعا في عدة مجالات كالعلوم الاكتوارية والسـكانية، والعلوم الطبية والبيولوجية، والرياضـيات والهندسـة وعلوم الحاسـب، وعلوم الفضـاء وغيرها.

2. هدف وحدود البحث:

تهدف هذه الدراســـة إلى تقدير القيم البينية المفقودة في بيانات الوفاة باســـتخدام مجموعة من الطرق الرياضـية والإحصـائية المعلمية واللامعلمية (الاسـتكمال الخطي Linear interpolation - الاستكمال بكثيرات الحدود Polynomial interpolation - الشرائح التكعيبية Cubic splines - الاستكمال التكعيبي المتدرج لهيرمت Hermite باستخدام كثيرات الحدود (PCHIP) Piecewise cubic Hermite interpolating polynomial - شـــرائح التمهيـد المُقطَّعـة Discretized smoothing splines) والمقارنة بينها.

ولأســـباب تتعلق بتوفر البيانات ومصـــداقيتها، وحتى تكون المقارنة بين الطرق مبنية على أســاس موثوق، فقد تم تطبيق الأســاليب الســابقة على بيانات كل من أعداد الوفيات الخام ومعدلات الوفاة لبعض الدول الأوروبية عن فترة ثلاث ســنوات من عام 2018 إلى عام 2020 (Human Mortality Database, 2021).

3. أهمية البحث:

تتمثل أهمية هذه الدراسة في تقديم بعض الطرق الإحصائية للاستكمال التي لم يتم تناولها أو تطبيقها لهذا الغرض على بيانات ديموجرافية أو على معدلات الوفاة في الدراسات باللغة العربية (في حدود علم الباحث)، والمقارنة بينها وبين بعض الطرق الشائعة، وكذلك المساهمة في إثراء المكتبة العربية في مجال التخصص.

4. الدراسات السابقة:

هنالك العديد من الدراسات التي تناولت تقدير القيم المفقودة بأساليب ونماذج إحصائية متعددة، وبالتطبيق على شتى المجالات. منها ما هو في مجالات الجيولوجيا والرياضيات والهندسة وغيرها، وأخرى في المجال الصحي أو الطبي، ومنها ما هو على بيانات ديموجرافية أو بيانات الوفيات سواء أعداد الوفيات أو معدلات الوفاة.

فقد تناولت دراسة (Garcia, D., 2010) تطبيق شرائح التمهيد المقطعة، وأساسها الشرائح الجزائية، لتمهيد البيانات المنفصلة في حالة وجود قيم مفقودة، وذلك بالتطبيق على بيانات المتوسطات السنوية لدرجات حرارة سطح الأرض في مواقع مختلفة، المنشورة بواسطة مكتب الأرصاد الجوية بالمملكة المتحدة. وتم تطبيق الأسلوب المقترح في حالتي السلاسل الزمنية ذات متغير واحد وذات متغيرين. وتميز أسلوب الشرائح المقترح بوجود معلمات يمكن التحكم فيها لضبط القيم الممهدة في حالة وجود قيم مفقودة أو متطرفة أو كلاهما معا في البيانات. وتناولت دراسة (محمد حبيب & حافظ محمد، 2011) مقارنة بين بعض نماذج الانحدار اللامعلمية الشائعة الاستخدام، وهي مقدر Nadaraya-Watson ومقدر الانحدار الخطي الموضعي Local Linear Estimator (وكلاهما يعتمد على تقديرات Kernel، وشرائح التمهيد Smoothing Splines، وانحدار الشرائح الجزائية Penalized Spline Regression، حيث تم تطبيق الطرق الأربعة على بيانات محاكاة لثلاث نماذج رياضية، وقد توصلت الدراسة إلى أن طرق الشرائح كانت أفضل من طرق kernel في نمذجة البيانات الناتجة عن المحاكاة. وناقشت دراسة (Azizan, I., et al., 2018) تطبيق نوعين من الشرائح التكعيبية (الطبيعية، العقد الحدودية النهائية) لنمذجة بيانات هطول الأمطار والتنبؤ بها باستخدام بيانات الأمطار الشهرية المنشورة عن إدارة الأرصاد الجوية في ماليزيا عامي 2014 و2015. وتم تقدير القيم السالبة والمفقودة على منحنى الاستكمال في بعض الفترات الفرعية باستخدام طريقة PCHIP. بينما قارنت دراسة (Rabbath, C. A., Corriveau, D., 2019) بين طرق الاستكمال الخطي والشرائح التكعيبية وPCHIP في مجال الديناميكا الهوائية كطرق لتقدير منحنى مسار الحركة لقذائف الأسلحة الصغيرة، حيث تم عمل محاكاة لستة مسارات محتملة وتطبيق الطرق الثلاثة على بيانات المحاكاة. وقد أعطت طريقة PCHIP أقرب النتائج لوصف الديناميكا الهوائية للقذيفة. وقدمت دراسة (Zaghiyan, M. R. et al., 2021) تطبيقا لعدد من طرق الاستكمال اشتملت على الاستكمال الخطي وطريقة أقرب جار والشرائح التكعيبية وPCHIP لغرض استكمال بيانات مستويات المياه الجوفية خلال فترات القياس الدورية غير المنتظمة، حيث استخدمت قياسات مستويات المياه لعدد 46 بئراً في إحدى مناطق إيران على مدار 20 عامًا. وتمت

المقارنة بين الطرق المستخدمة بطريقة المصادقة المتقاطعة المعممة (GCV) وذلك في ثلاث حالات لمستويات المياه، المنخفض والمتوسط والعالي. وأثبتت النتائج أن طريقة PCHIP كانت الأكثر دقة يليها الشرائح التكعيبية ثم الاستكمال الخطي.

وفي المجال الصحي، تناولت دراسة (Bazo-Alvarez, et al., 2020) معالجة القيم المفقودة في تحليل السلاسل الزمنية المقطوعة للبيانات الطولية باستخدام كل من الانحدار المجزأ (segmented regression) والنماذج المختلطة، وذلك بالتطبيق على بيانات السجلات الصحية للمرضى الذين يتناولون مضادات الذهان بالمملكة المتحدة، وأوضحت النتائج أن تقديرات الانحدار المجزأ تكون غير متحيزة عندما تكون البيانات مفقودة بشكل عشوائي Missing at Random (MAR)، كما أن استخدام النماذج المختلطة كان من شأنه تحسين دقة التقدير في حالات محددة. كما قامت دراسة (Acal, C., et al., 2021) بتطبيق نموذج الانحدار function-on-function regression model لتقدير القيم المفقودة والتنبؤ بعدد الحالات غير المرصودة الخاصة بوباء كورونا COVID-19 (الإصابات الجديدة والوفيات والمتعافون) في المستشفيات ووحدات العناية المركزة في إسبانيا. واهتمت دراسة (Simos, T. E., et al., 2021) بتقدير رقم التكاثر الأساسي R_0 لتفشي فيروس COVID-19 باستخدام نموذج (SIR) Susceptible, Infectious, or Recovered، وهو أحد النماذج المعروفة لتقدير حالات الإصابة والتعافي عند انتشار الأوبئة. وقد توصلت الدراسة إلى صيغة مباشرة لحساب R_0 بناء على البيانات الفعلية، ثم تم تطبيق كل من نماذج الفروق المحدودة والشرائح التكعيبية وPCHIP والمربعات الصغرى الخطية لتقدير واستكمال قيم R_0. ورغم تقارب نتائج الطرق المستخدمة غير أن طريقة المربعات الصغرى فقط هي التي كانت تقديراتها لقيمة R_0 دائما موجبة، ولذلك تم الاعتماد على نتائجها للتنبؤ بدرجة انتشار الوباء خلال فترة التقدير.

وقامت دراسة (Currie, Durban, & Eilers, 2004) باستخدام طريقة الشرائح الجزائية P-splines لتمهيد معدلات الوفاة والتنبؤ بها. واستخدمت نموذجاً خطياً معمماً جزائياً بالتطبيق على مجموعتين من البيانات (السكان وأصحاب المعاشات) لوفيات الذكور في المملكة المتحدة خلال الفترة من 1947 حتى 1999. وقدمت دراسة (Andreopoulos, P., et al., 2019) توزيعا جديدا يدمج بين أشكال دوال توزيعات جومبيرتز وماكيهام ودالة توزيع بيتا. وتم التطبيق على بيانات معدلات الوفيات الخاصة بالعمر والجنس في اليونان لعام 2011. وقد أظهر التوزيع المقترح قدرة تنبؤية أعلى للذكور والإناث، خاصة في الأعمار الكبيرة، مقارنة بالشكل التقليدي لتوزيعات جومبيرتز- ماكيهام. وتناولت دراسة (McNeil, N., et al., 2011) استخدام الشرائح التكعيبية الطبيعية في استكمال البيانات الديموغرافية المتمثلة في عدد السكان وعدد الوفيات ومعدلات الخصوبة العمرية في إيطاليا، مع ضبط النموذج بما يحقق الخصائص المرغوبة لوظيفة الاستكمال دون النظر إلى زيادة درجة التمهيد. وأخيرا، فقد اقترحت دراسة (Schmertmann, C., 2021) نوعا معدلا من شرائح التمهيد الجزائية P-splines أُطلق عليه D-splines. وتم تطبيق نوعي الشرائح الأصلية والمعدلة لتقدير معدلات الوفاة، مع تطبيق ثلاث حالات للنموذج المعدل على

مجتمعين سكانيين مختلفي الحجم. وأظهرت النتائج أن النموذج المقترح يُنتج تقديرات أفضل من حيث الدقة والتحيز في حالة العينات الصغيرة.

ومما سبق، يتضح أن بعض أنواع شرائح التمهيد، خاصة تلك المشتقة من الشرائح التكعيبية والشرائح الجزائية، بالإضافة إلى استخدامها في تمهيد البيانات، يمكن أن تستخدم لغرض استكمال البيانات أو تقدير القيم المفقودة بها، وفي تطبيقات متعددة. كما تجدر الإشارة إلى أنه لم يتم التطرق إلى الدراسات الخاصة باستكمال جداول الحياة المختصرة، والتي يختلف فيها مفهوم الاستكمال وأساليبه ووظيفته عن حالة القيم المفقودة (الأشقر، ا.، زايد، م.، 2020).

5. منهجية البحث:

تعتمد هذه الدراسة على منهج تحليلي للمقارنة بين بعض الطرق الرياضية والإحصائية لتقدير (أو استكمال) القيم المفقودة في بيانات الوفيات.

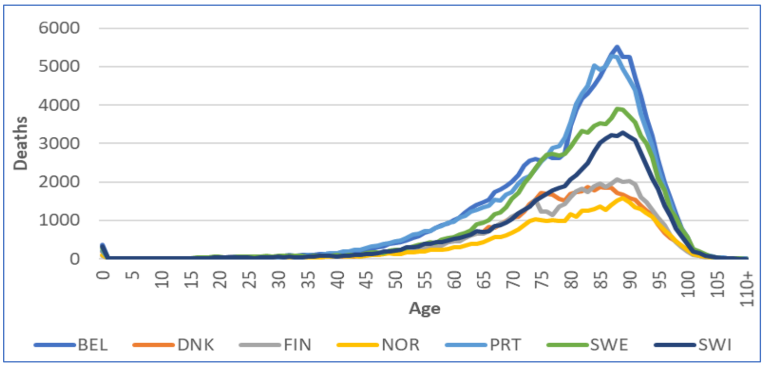
1.5 عينة الدراسة (بيانات الوفيات)

تم تطبيق هذه الدراسة على بيانات الوفيات المنشورة في قاعدة بيانات HMD (Human Mortality Database, 2021)، والمتمثلة في كل من أعداد الوفيات ومعدلات الوفاة الخام. واشتملت عينة الدراسة على بيانات الوفيات لسبع دول (بلجيكا BEL - الدنمارك DNK - فنلندا FIN - النرويج NOR - البرتغال PRT - السويد SWE - سويسرا SWI) عن الفترة من 2018 حتى نهاية عام 2020. ويوضح الجدولان (1) و (2) الإحصاءات الوصفية الأساسية لبيانات الدراسة، والشكلان (1) و (2) بيانات الوفيات عن عام 2020.

جدول رقم (1): الإحصاءات الوصفية لأعداد الوفيات

		BEL	DNK	FIN	NOR	PRT	SWE	SWI
2020	Min	1	0	0	1	2	1	0
	Max	5522	1879	2074	1590	5273	3897	3275
	Av	1143	492	499	366	1111	884	686
	Sd	1557.4	633.2	649.4	470.3	1524.0	1214.2	959.0
	CV (%)	136.2	128.6	130.2	128.6	137.1	137.4	139.7
2019	Min	1	1	0	0	1	1	0
	Max	4696	1816	2096	1594	4802	3476	2846
	Av	980	486	486	367	1007	800	611
	Sd	1302.8	615.9	634.8	470.9	1365.7	1079.0	828.9
	CV (%)	132.9	126.7	130.6	128.5	135.6	134.9	135.7
2018	Min	1	0	0	0	1	0	1
	Max	4700	1846	2131	1591	4735	3625	2779
	Av	997	498	491	368	1018	830	604
	Sd	1327.6	627.8	638.8	474.4	1388.7	1119.7	818.9
	CV (%)	133.1	126.2	130.0	128.9	136.3	134.8	135.5

ومن الجدول الســـابق، يمكن ملاحظة أن هناك بعض الاختلاف بين الدول من حيث أعداد الوفيات (المتوســطات وأكبر قيمة لكل حالة)، وإن كان لا يوجد اختلاف مؤثر من حيث مدى تجانس أو تشتت القيم (معاملات الاختلاف)، كما يمكن استنتاج أن نسبة الزيادة في أعداد الوفيات عام 2020 كانت أكبر من تلك الخاصة بعام 2019.

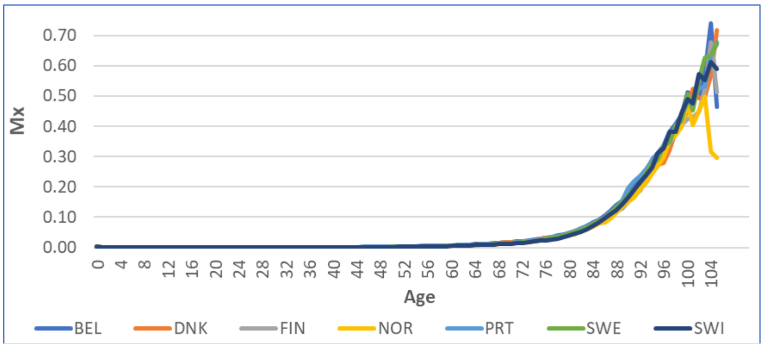


الشكل رقم (1): أعداد الوفيات لعام 2020

جدول رقم (2): الإحصاءات الوصفية لمعدلات الوفاة الخام

		BEL	DNK	FIN	NOR	PRT	SWE	SWI
2020	Min	0.00004	0.00002	0.00000	0.00002	0.00004	0.00002	0.00005
	Max	0.73927	0.71781	0.67739	0.49751	0.67919	0.67273	0.61179
	Av	0.07303	0.06852	0.06539	0.05867	0.07272	0.07172	0.06962
	Sd	0.150	0.144	0.137	0.119	0.151	0.153	0.148
	CV (%)	205.7	210.1	209.6	203.3	207.2	212.9	212.5
2019	Min	0.00004	0.00003	0.00002	0.00005	0.00005	0.00003	0.00002
	Max	0.57660	0.62954	0.58000	0.99999	0.74158	0.64264	0.73522
	Av	0.06371	0.06769	0.06461	0.06807	0.06634	0.06473	0.06417
	Sd	0.132	0.140	0.134	0.164	0.139	0.137	0.141
	CV (%)	207.3	207.0	207.1	241.3	210.1	212.0	219.3
2018	Min	0.00006	0.00002	0.00000	0.00000	0.00005	0.00001	0.00005
	Max	0.79186	0.58231	0.59925	0.90452	0.57380	0.63778	0.57043
	Av	0.07070	0.07016	0.06680	0.07037	0.06607	0.06879	0.06218
	Sd	0.153	0.143	0.138	0.159	0.135	0.146	0.132
	CV (%)	217.0	204.3	206.1	226.3	203.9	212.6	212.1

ومن الجدول الســـابق، يمكن ملاحظة أنه لا يوجد اختلاف مؤثر بين الدول الســبع من حيث مدى تجانس أو تشـــتت قيم معدلات الوفاة، كما يظهر أن معدلات الوفاة قد ارتفعت، في المتوســـط، بنسبة بسيطة عام 2020 لجميع الدول فيما عدا النرويج (NOR).



الشكل رقم (2): معدلات الوفاة الخام لعام 2020

2.5 أساليب استكمال وتقدير بيانات الوفيات

تمثلت الأساليب المستخدمة في هذه الدراسة فيما يلي:

1.2.5 الاستكمال الخطي Linear Interpolation

يعتبر الاستكمال الخطي من الطرق الأساسية للاستكمال وأبسطها. وفيه يتم تقدير أي نقطة تقع بين نقطتين معرَّفتين على أساس ربط هاتين النقطتين بخط مستقيم، وافتراض أن النقطة المقدرة تقع على هذا الخط المستقيم.

وفي حالة البيانات ثنائية الأبعاد، إذا كانت قيم x مرتبة ترتيبا تصاعدياً، أي $x_i < x_{i+1}$ ، وكانت $x_i < x < x_{i+1}$ ، فإن الاستكمال الخطي عند x بمعلومية النقطتين $(x_i, y_i), (x_{i+1}, y_{i+1})$ هو :(Siauw, T., & Bayen, A., 2015)

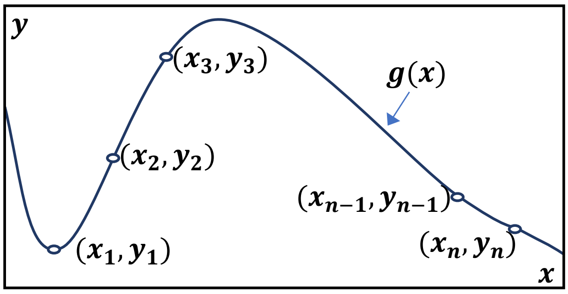
$$\hat{y}(x) = y_i + \frac{(y_{i+1} - y_i)(x - x_i)}{(x_{i+1} - x_i)} \quad (1)$$

وغالبا يستخدم الاستكمال الخطي في حالة عدم وجود معلومات محددة حول شكل توزيع البيانات ومدى تجانسها أو تباينها، وخاصة إذا كانت نقاط البيانات قريبة من بعضها البعض.

2.2.5 الاستكمال بكثيرات الحدود Polynomial Interpolation

يمكن النظر إلى الاستكمال بكثيرات الحدود على أنه الحالة العامة للاستكمال الخطي الذي يعتبر كثيرة حدود من الدرجة الأولى. فبدلا من إيجاد معادلة الخط المستقيم الواصل بين نقطتين معلومتين في حالة الاستكمال الخطي، يتم إيجاد كثيرة الحدود من درجة محددة التي تمر عبر عدد معلوم من النقاط.

نفترض أن لدينا $n+1$ من النقاط $\{(x_0, y_0), (x_1, y_1), \ldots, (x_n, y_n)\}$ ، ونريد إيجاد كثيرة الحدود $g(x)$ من الدرجة d التي تمر عبر تلك النقاط، كما هو موضح في الشكل رقم (3) أدناه. :(Meseguer, A., 2020)



الشكل رقم (3)

فإذا كانت كثيرة الحدود هي:

$$g(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta_d x^d, \qquad (2)$$

حيث θ_k $(k = 0,1,2,\dots,d)$ هي $d+1$ من المعاملات المجهولة.

وكما يتم في حالة الخط المستقيم، يمكن حساب المعاملات θ_k بفرض أن كثيرة الحدود يجب أن تمر عبر النقاط المحددة (x_i, y_i)، أي أن:

$$g(x_i) = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \dots + \theta_d x_i^d = y_i, \quad (i = 1,2,\dots,n) \qquad (3)$$

ويمكن التعبير عن (3) بالنظام الخطي العام التالي:

$$\begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^d \\ 1 & x_2 & x_2^2 & \dots & x_2^d \\ 1 & x_3 & x_3^2 & \dots & x_3^d \\ . & . & . & & . \\ . & . & . & & . \\ . & . & . & & . \\ 1 & x_n & x_n^2 & \dots & x_n^d \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ . \\ . \\ . \\ \theta_d \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_n \end{bmatrix} \qquad (4)$$

ويشار عادةً للمصفوفة المربعة التي تظهر على الجانب الأيسر من (4) بمصفوفة Vandermonde's matrix [1] $(V_{n\times(d+1)})$ حيث $V_{ij} = x_i^{j-1}$، ويكون:

$$\Delta V = \prod_{0 \le i \le j \le n} (x_j - x_i) \qquad (5)$$

ولأن الإحداثيات x_iمختلفة، فإن $\Delta V \neq 0$، وبالتالي، فإن للمعادلة (4) دائماً حل وحيد، ولذلك، تكون كثيرة الحدود $g(x)$ الناتجة وحيدة أيضاً.

١ - Alexandre The´ ophile Vandermonde (١٧٣٥-١٧٩٦)، عالم رياضيات فرنسي اشتهر بمساهماته في نظرية المحددات.

3.2.5 الشرائح التكعيبية Cubic Splines

تعتبر الشـــرائح التكعيبية cubic splines من أشـــهر أنواع ما يعرف بشـــرائح التمهيد (smoothing splines)، والتي يمكن تعريفها على أنها دالة كثيرة حدود $S(x)$ من درجة معينة d وممهَّدة بحيث تكون قابلة للاشـــتقاق أو يمكن تقدير جميع مشـــتقاتها حتى الرتبة (d-1)، يمكن التعبير عنها كما يلي:

$$S(x)=\sum_{j=0}^{d}\theta_j x^j+\sum_{r=d+1}^{d+k}\theta_r(x-\tau_{r-d})^d \qquad (6)$$

حيث:

$\theta_{...}$: مجموعة من المعاملات المجهولة.

$\tau_{K\ (K=1,2,...,k)}$: مجموعة من العقد المتتالية $(\tau_1<\tau_2<...<\tau_k)$ في نطاق الدالة $S(x)$.

وفي حالة الشـــرائح التكعيبية يتم توفيق كثيرة حدود متعددة التعريف من الدرجة الثالثة في الفترة بين كل عقدتين متتاليتين.

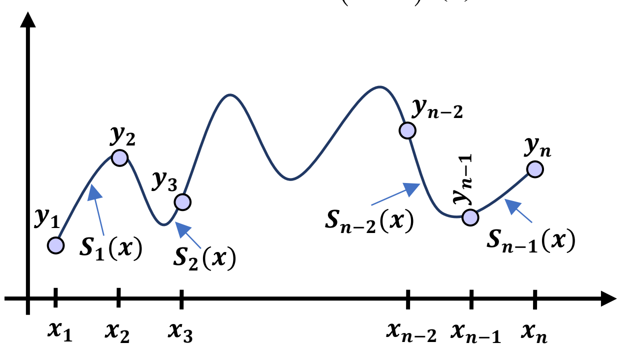
وعند الاســتكمال باســتخدام الشــرائح التكعيبية، نفترض أن أي نقطتين متتاليتين (x_i, y_i) و (x_{i+1}, y_{i+1}) ، حيث $x_i \le x \le x_{i+1}$ ، ترتبطـان بكثيرة حدود من الـدرجـة الثـالثـة تكون على الصورة التالية لعدد n من النقاط:

$$S_i(x_i)=\sum_{j=0}^{3}\theta_{ij}x_i^{\,j} \quad ,i=1,......,n-1 \qquad (7)$$

ولإيجاد المعاملات θ_{ij} لكل دالة من الدوال التكعيبية $(n-1)$ دالة (الشـــكل رقم (4)) لكل منها أربع معـاملات)، أي $4(n-1)$ من المعـاملات المجهولـة، نحتـاج إلى $4(n-1)$معـادلـة (Siauw, T., & Bayen, A., 2015)، ويكون:

$$S_i(x_i)=y_i \quad , \quad S_i(x_{i+1})=y_{i+1} \qquad (8)$$

وينتج عن الصيغة السابقة (9) $2(n-1)$ من المعادلات.



الشكل رقم (4)

ثم للوصول إلى دالة ممهدة يتم فرض قيود بأن يكون لها مشتقتان أولى وثانية متصلتان عند نقاط البيانات $i = 2,...,n-1$.

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}), \quad i = 1,...,n-2,$$
$$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}), \quad i = 1,...,n-2, \tag{9}$$

والتي تعطينا $2(n-1)$ من المعادلات.

وأخيرا، هناك حاجة إلى معادلتين إضافيتين لحساب معاملات $S_i(x_i)$، حيث يتم عادة افتراض أن المشتقات الثانية تساوي الصفر عند نقاط النهاية، أي أن:

$$S_1''(x_1) = 0 \;,\; S_{n-1}''(x_n) = 0 \tag{10}$$

وبحل المعادلات الخطية الناتجة عن الصيغ (8) و(9) و(10) نصل إلى معادلات الشرائح التكعيبية.

4.2.5 الاستكمال التكعيبي المتدرج لهيرمت باستخدام كثيرات الحدود (PCHIP) Piecewise cubic Hermite interpolating polynomial

تعتبر هذه الطريقة حالة خاصة من شرائح التمهيد التكعيبية، حيث يتم تقدير كثيرة حدود مجزأة أو متدرجة (piecewise) من الدرجة الثالثة لها شكل صيغة استكمال Hermite[٢] بين كل نقطتين، ووفقا لهذه الصيغة يجب أن يحقق الاستكمال شرطين هما:

- أن تكون القيم المستكملة عند نقاط البيانات المعلومة مطابقة لقيم هذه البيانات، أي $g(x_j) = f_j$.
- أن تتساوى قيمة المشتقة الأولى لدالة الاستكمال مع قيمة المشتقة الأولى الفعلية (أو التي يتم تقديرها من البيانات الفعلية) عند نقاط البيانات المعلومة، أي $g'(x_j) = f'_j$.

وتكون شريحة التمهيد الناتجة متصلة ولها مشتقة أولى متصلة عند جميع نقاط البيانات الفعلية.

ويعتبر استكمال لاجرانج حالة خاصة من استكمال Hermite في حالة عدم وجود قيود توجب تساوي المشتقة الأولى لدالة الاستكمال مع المشتقة الأولى لدالة البيانات، كما يمكن تمييز الخصائص التالية لطريقة PCHIP مقارنة بالشرائح التكعيبية (Rabbath, A., & Corriveau, D., 2019):

- اشتراط تطابق قيم المشتقات من الدرجة الأولى، وليس فقط الاكتفاء بوجود تلك المشتقات، عند نقاط البيانات المعلومة.

---
[٢] ترجع هذه التسمية إلى عالم الرياضيات الفرنسي Charles Hermite (1822 - 1901).

- الاستكمال بطريقة PCHIP أقل نعومة لكنه أكثر دقة من الشرائح التكعيبية.
- تكون فرص ظهور انحرافات كبيرة أو قيم متطرفة عند تطبيق PCHIP أقل منها في حالة الشرائح التكعيبية.

ولإيجاد كثيرات حدود Hermite التكعيبية، يتم حل أنظمة المعادلات التالية:

$$g(x_j) = \theta_0 + \theta_1 x_j + \theta_2 x_j^2 + \theta_3 x_j^3 = f_j,$$
$$g'(x_j) = \theta_1 + 2\theta_2 x_j + 3\theta_3 x_j^2 = f_j', \quad (j = 0, 1, 2, ..., n) \tag{11}$$

وبالتالي، لأي نقطتين $(x_0, f_0), (x_1, f_1)$، تكون قيم معاملات دالة الاستكمال هي:

$$\theta_0 = f_0,$$
$$\theta_1 = f_0',$$
$$\theta_2 = 3f_1 - 3f_0 - f_1' - 2f_0',$$
$$\theta_3 = -2f_1 + 2f_0 + f_0' + f_1' \tag{12}$$

5.2.5 شرائح التمهيد المُقطَّعة Discretized smoothing splines

تعتبر شرائح التمهيد المقطعة حالة خاصة من الشرائح الجزائية تم اقتراحها بواسطة Garcia (Garcia, 2010)، وتقوم على الانحدار بطريقة المربعات الصغرى الجزائية مع استخدام تحويلة جيب التمام المنفصل Discrete Cosine Transform (DCT). وتتميز هذه الطريقة بأنها لا تتأثر بوجود قيم مفقودة في البيانات، ولذلك يعتبر استخدامها مناسبا عند تسوية أو تمهيد البيانات التي تحتوي على قيم مفقودة. وفيما يلي عرض موجز لهذه الطريقة.

لنأخذ بعين الاعتبار النموذج التالي:

$$y = \hat{y} + \varepsilon, \tag{13}$$

حيث y : القيم الفعلية، ε : الخطأ العشوائي بمتوسط يساوي الصفر وتباين غير معلوم، و $\hat{y}$ القيم الممهدة.

ووفقا لانحدار المربعات الصغرى الجزائية (Wahba, 1990)، يتم تدنية المقدار:

$$F(\hat{y}) = RSS + \lambda P(\hat{y}) = \|y - \hat{y}\|^2 + \lambda P(\hat{y}), \tag{14}$$

حيث:

RSS : مجموع مربعات الخطأ العشوائي (وهو مقياس لمدى اقتراب القيم الممهدة من القيم الفعلية).

P : حد الجزاء في النموذج (ويعبر عن مدى سلاسة القيم الممهدة).

λ : تمثل معلمة التمهيد (كلما زادت قيمة λ كلما كانت القيم الممهدة أكثر نعومة).

$\| \|$: تشير إلى القاعدة الإقليدية Euclidean norm.

ولتحويل النموذج السابق إلى شكل شريحة تمهيد smoothing spline، يتم التعبير عن حد الجزاء في النموذج بدلالة المشتقات العليا للقيم الممهدة $\hat{y}$ كالتالي:

$$P\left(\hat{y}\right)=\left\|D\,\hat{y}\right\|^{2}, \qquad (15)$$

حيث D تمثل مصفوفة مثلثية القطر tridiagonal معرفة كالتالي:

$$D_{i,i-1}=\frac{2}{h_{i-1}\left(h_{i-1}+h_{i}\right)},D_{i,i}=\frac{-2}{h_{i-1}h_{i}},D_{i-1,i}=\frac{2}{h_{i}\left(h_{i-1}+h_{i}\right)} \qquad (16)$$

و h_i هي المسافة أو الفرق بين $\hat{y}_i$ و $\hat{y}_{i+1}$.

وينتج عن عملية تدنية المقدار $F\left(\hat{y}\right)$ النظام الخطي التالي:

$$(I_n+\lambda D^{T}D)\hat{y}=y \qquad (17)$$

ويتم تقدير معلمة التمهيد بطريقة التحقق المتقاطع المعمم generalized cross validation (GCV) (Wahba, 1990)، كالتالي (Garcia, 2010):

$$\lambda=\arg\min(GCV),\ \ GCV(\lambda)=\frac{n\sum_{i=1}^{n}\left(\hat{y}_i-y_i\right)^{2}}{\left(n-\sum_{i=1}^{n}\left(1+\lambda\gamma_i^{2}\right)^{-1}\right)^{2}}, \qquad (18)$$

حيث: $\left(\gamma_i^{2}\right)_{i=1,.....,n}$ هي القيم أو الجذور الكامنة لـ $D^{T}D$.

وعند استخدام تستخدم تحويلة جيب التمام المنفصل (DCT)، يكون:

$$GCV(\lambda)=\frac{n\sum_{i=1}^{n}\left(\frac{1}{1+\lambda\gamma_i^{2}}-1\right)^{2}DCT_i^{2}(y)}{\left(n-\sum_{i=1}^{n}\frac{1}{1+\lambda\gamma_i^{2}}\right)^{2}} \qquad (19)$$

ثم يتم الحصول على القيم الممهدة $\hat{y}$ كما يلي:

$$\hat{y}=U\,\Gamma\,\mathrm{DCT}(y)=\mathrm{IDCT}\left(\Gamma\,\mathrm{DCT}(y)\right), \qquad (20)$$

حيث تشير IDCT إلى تحويل جيب التمام المنفصل العكسي (Garcia, 2010).

3.5 اختبارات جودة التوفيق

للحكم على مدى اقتراب القيم المقدرة من القيم الأصلية، سيتم الاعتماد على المقاييس التالية:

1.3.5 جذر متوسط مربعات البواقي Root Mean Square Error

وفقا لهذا المقياس، يتم حساب جودة توفيق القيم المقدرة من حيث مدى اقترابها من القيم الفعلية بالاعتماد على مجموع مربعات الخطأ العشوائي $\text{SSE} = \sum_x \left(y_x - \hat{y}_x\right)^2$ ثم إيجاد القيمة:

$$\text{RMSE} = \sqrt{\frac{\sum_x \left(y_x - \hat{y}_x\right)^2}{n}} \qquad (21)$$

حيث تمثل y_x القيمة الفعلية، و$\hat{y}_x$ القيمة المتوقعة. وكلما كان المقدار السابق أقل كلما كان التقدير أكثر دقة.

2.3.5 اختبار مربع كاي (χ^2) The Chi-Square Test

يتم في هذا الاختبار حساب القيمة χ^2 كالتالي:

$$\chi^2 = \sum \frac{\left(y_x - \hat{y}_x\right)^2}{\hat{y}_x} \qquad (22)$$

ثم تقارن القيمة المحسوبة بقيمة تستخرج من جدول توزيع χ^2 عند مستوى الدلالة المستخدم وبدرجات حرية محددة، فإذا كانت أصغر من القيمة الجدولية يكون التقدير مقبولا.

6. التحليل والنتائج:

لغرض المقارنة بين الأساليب المستخدمة في هذه الدراسة من حيث دقة تقدير القيم المفقودة، فقد تم التطبيق على بيانات الوفيات، لجميع الدول ولجميع السنوات، بعد حذف مجموعة من القيم على اعتبار أنها قيم مفقودة. وبالتالي يتركز الاهتمام على مقارنة الأساليب المطبقة من حيث دقة التقدير فقط، ولا يدخل في نطاق اهتمام هذه الدراسة محاولة تصنيف القيم المفقودة حسب سبب الفقد، سواء اعتبرت تلك القيم مفقودة عشوائياً بشكل تامّ missing completely at random (MCAR) أو مفقودة عشوائياً missing at random (MAR) أو غير مفقودة عشوائياً missing not at random (MNAR)، وذلك لأنه يمكن تطبيق الأساليب المقترحة في هذه الدراسة في تقدير القيم المفقودة بغض النظر عن تصنيفها حسب السبب.

وبناء على ما سبق، ولتحقيق الهدف من هذه الدراسة، فقد تم تطبيق الأساليب المقترحة لتقدير القيم المفقودة، سواء على أعداد الوفيات أو معدلات الوفاة، في حالتين:

- الحالة الأولى: وجود قيمة واحدة أو مجموعة متفرقة من القيم المفقودة عند أعمار مختلفة.
- الحالة الثانية: وجود سلسلة متصلة من القيم المفقودة عند أعمار مختلفة.

وفي كل من الحالتين الأولى والثانية، وحتى تكون المقارنة بين الأساليب المستخدمة أكثر تفصيلا، تم اختيار الأعمار التي تناظر القيم المفقودة (أ) مرة بحيث تشمل مجموعة من الأعمار التي لا توجد عندها تقلبات ملحوظة في منحنى البيانات، و (ب) مرة أخرى بحيث تشمل أعمارا يغلب عليها عكس ذلك. ولاختيار تلك الأعمار، تم الاستناد إلى قياس تغير إشارات الفروق بين القيم المتتابعة سواء لأعداد الوفيات أو لمعدلات الوفاة.

وقد تم تقدير القيم المفقودة وتقدير معلمات النماذج الإحصائية، وإجراء اختبارات جودة التوفيق بالاستعانة ببرنامجي MATLAB-(R2021b) و MS-Excel. وتمثلت خطوات التطبيق فيما يلي:

١. تقدير القيم المفقودة في أعداد الوفيات ومعدلات الوفاة عن طريق تطبيق الصيغ الرياضية للاستكمال (المعادلة رقم (1)) وتوفيق النماذج المستخدمة، وذلك لكل حالة من حالات القيم المفقودة ولكل دولة ولكل سنة على حدة.
٢. حساب جذر متوسط مربعات الخطأ العشوائي (RMSE) للقيم المقدرة (المعادلة رقم (21))، وإجراء اختبار χ^2 لجودة التوفيق (المعادلة رقم (22)).

ونستعرض فيما يلي ملخص نتائج التحليل.

1.6 تقدير القيم المفقودة بالتطبيق على أعداد الوفيات

يوضح الجدولان (3) و (4) ملخص نتائج اختبارات جودة التوفيق عند تقدير القيم المفقودة في أعداد الوفيات، وتشمل قيمة RMSE لكل أسلوب (الاستكمال الخطي (Lin) - الاستكمال بكثيرات الحدود (Pol) - الشرائح التكعيبية (CS) - الاستكمال التكعيبي المتدرج لهيرمت باستخدام كثيرات الحدود (PCHIP) - شرائح التمهيد المُقطَّعة (DSS))، بالإضافة إلى القيمة المحسوبة لاختبار χ^2 عند أفضل تقدير.

جدول رقم (3): جذر متوسط مربعات الخطأ (RMSE)
تقدير القيم المفقودة لأعداد الوفيات - الحالة الأولى

السنة	الدولة	الحالة الأولى - (أ)						الحالة الأولى - (ب)					
		Lin	Pol	PCHIP	CS	DSS	χ^2	Lin	Pol	PCHIP	CS	DSS	χ^2
2020	BEL	115.9	520.0	110.3	117.8	112.8	24.5*	6.9	241.4	6.6	19.0	6.0	7.8
	DEN	39.5	102.0	32.3	39.8	34.8	8.4	7.7	41.1	9.5	31.8	10.5	11.4
	FIN	69.0	177.4	79.2	96.4	71.1	24.6*	39.3	21.9	39.2	26.9	45.2	91.3*
	NOR	11.3	102.2	10.6	11.6	12.1	5.9	45.0	43.2	44.1	41.1	47.5	23*
	PRT	22.4	382.2	27.8	39.3	29.3	6.9	38.5	258.1	38.5	11.8	72.9	9.7
	SWE	17.1	236.2	24.0	27.8	20.9	4.9	34.0	59.3	29.3	13.6	43.0	6.2
	SWI	7.3	262.5	17.8	35.1	23.6	4.1	7.9	148.3	12.2	16.4	10.3	7.3
2019	BEL	42.2	339.6	39.7	43.0	78.8	16.9	70.0	427.3	82.9	84.5	94.9	20.1*
	DEN	65.4	97.9	66.5	71.2	59.2	26.7*	38.8	40.6	49.6	66.4	44.6	21.5*
	FIN	52.1	181.7	53.7	114.2	26.7	6.1	70.2	228.7	78.0	97.9	60.6	20.6*
	NOR	40.7	85.2	41.6	54.8	33.7	10.9	58.9	70.4	63.0	74.7	58.9	24.7*
	PRT	112.2	353.9	111.4	118.8	101.6	24.6*	55.9	294.3	58.6	84.7	65.0	12
	SWE	67.2	205.5	67.8	77.2	52.1	6.4	50.4	140.7	70.6	94.1	71.2	11.4
	SWI	18.8	179.3	14.0	20.9	12.6	13.1	22.3	205.6	15.4	24.4	12.7	3
2018	BEL	34.2	355.4	36.4	44.7	29.7	7.8	8.2	101.6	8.2	10.1	34.8	6.6
	DEN	31.3	102.3	19.6	18.7	28.4	1.6	14.5	26.9	15.8	16.5	29.3	10.2
	FIN	61.2	207.5	53.9	60.1	60.6	21*	9.0	33.7	10.3	10.5	8.6	7.7
	NOR	38.0	85.5	38.5	35.9	38.4	18.8*	4.3	15.7	5.3	7.3	3.0	1.2
	PRT	41.2	326.1	79.5	79.0	35.1	3.6	9.4	142.1	12.3	15.4	9.1	40.9*
	SWE	86.9	259.3	70.7	77.7	74.8	17.1*	11.2	44.7	12.6	13.3	11.1	12.8
	SWI	14.7	193.1	17.1	27.9	20.2	3.9	9.3	63.8	9.5	10.5	8.8	9.3
		تم اختيار القيم المفقودة لتكون عند الأعمار: 90, 75, 60, 45, 30, 15 * الفروق بين القيم المقدرة والفعلية ذات دلالة إحصائية						تم اختيار القيم المفقودة لتكون عند الأعمار: 49, 37, 89, 76, 68, 31, 26, - (2018) 29, 25, 21, 7 (2020) 86, 36, 31, 23, 14, 9 - (2019) 23					

من الجدول الســابق، يتضــح أن الشــرائح الجزائية المقطعة والاســتكمال الخطي كانتا، في الغالب، أفضل طريقتين لتقدير القيم المفقودة المتقطعة في أعداد الوفيات.

جدول رقم (4): جذر متوسط مربعات الخطأ (RMSE)
تقدير القيم المفقودة لأعداد الوفيات - الحالة الثانية

السنة	الدولة	الحالة الثانية - (أ)						الحالة الثانية - (ب)					
		Lin	Pol	PCHIP	CS	DSS	χ^2	Lin	Pol	PCHIP	CS	DSS	χ^2
2020	BEL	99.6	655.3	95.0	114.5	122.3	20.3*	6.4	93.4	7.8	13.5	8.0	6.4
	DEN	31.8	50.4	43.8	39.7	40.4	14.6	5.5	23.7	6.1	7.2	5.8	9.5
	FIN	37.9	177.5	33.5	21.2	29.3	3.7	7.4	20.7	7.7	10.9	7.1	12.2
	NOR	68.0	150.8	60.9	28.3	74.8	15.4*	5.6	10.2	5.8	6.7	5.4	10.8
	PRT	89.2	527.1	55.9	55.8	77.8	16.3*	8.2	123.8	8.3	11.2	7.1	5.7
	SWE	118.0	307.6	112.1	71.4	90.2	10.5	9.6	57.0	9.6	14.1	9.8	15
	SWI	41.6	307.7	40.6	45.9	93.6	7.9	5.7	60.5	6.0	8.1	6.6	5.4
2019	BEL	104.0	587.5	74.2	72.1	181.8	29.6*	48.3	459.9	48.4	55.6	60.1	12
	DEN	37.7	73.0	40.1	48.6	36.9	12.9	23.1	61.8	26.7	34.3	25.6	8
	FIN	41.6	233.8	36.1	42.8	38.0	13.2	57.5	191.7	60.7	80.7	48.6	12.8
	NOR	45.5	125.0	44.0	44.5	48.9	8.8	23.2	71.0	27.1	35.5	26.8	4.2
	PRT	48.0	336.0	63.1	111.2	75.4	3.4	41.5	409.1	41.2	70.5	61.6	17.4*
	SWE	67.3	244.8	69.1	91.5	121.4	17.9	17.4	106.1	18.7	53.8	13.8	7.9
	SWI	77.5	296.2	72.7	47.6	98.8	12.7	39.2	123.8	42.5	59.9	52.7	11.7
2018	BEL	45.6	486.2	18.9	30.6	92.0	4.9	10.9	91.6	12.3	15.5	58.4	10.9
	DEN	22.8	108.3	36.1	53.5	59.6	3.3	6.7	24.6	7.4	9.4	23.1	13.5
	FIN	50.6	207.7	48.6	51.3	51.1	17.7*	9.3	28.6	9.3	9.3	9.6	15
	NOR	33.9	147.2	30.8	23.4	27.6	5.1	4.5	18.2	5.5	7.2	3.7	4.1
	PRT	113.1	418.3	100.3	101.4	78.4	19.1*	13.7	134.4	13.2	9.7	13.0	11.3
	SWE	30.0	282.3	28.1	112.0	71.9	6.4	13.4	40.1	13.9	11.1	14.9	12.7
	SWI	28.9	264.9	21.6	29.0	74.4	4.7	7.4	50.3	8.8	11.4	8.1	5.8
		تم اختيار القيم المفقودة لتكون عند الأعمار: 88, 89, 44, 45, 66, 67 * الفروق بين القيم المقدرة والفعلية ذات دلالة إحصائية						تم اختيار القيم المفقودة لتكون عند الأعمار: 36, 37, 17, 18, 25, 26 (2018) - 27, 30, 31, 76, 77, 26 (2019) - 8, 9, 22, 23, 30, 31 (2020)					

من الجدول السابق، يتضح أن أغلب الطرق المستخدمة، باستثناء الاستكمال بكثيرات الحدود، كانت مناسبة لتقدير القيم المفقودة المتتالية في أعداد الوفيات.

ومن النتائج السابقة، يمكن الاستدلال على أنه عند تقدير أو استكمال القيم المفقودة في بيانات أعداد الوفيات، كانت الشرائح الجزائية المقطعة بالإضافة إلى الاستكمال الخطي هما، في الغالب، أفضل طريقتين، مع بعض الاستثناءات التي كانت فيها طريقتي الشرائح التكعيبية وPCHIP هما الأفضل.

2.6 تقدير القيم المفقودة بالتطبيق على معدلات الوفاة

يوضح الجدولان (5) و (6) ملخص نتائج اختبارات جودة التوفيق عند تقدير القيم المفقودة في أعداد الوفيات، وتشمل قيمة RMSE لكل أسلوب بالإضافة إلى القيمة المحسوبة لاختبار χ^2 (لمعدلات الوفاة لكل ألف نسمة) عند أفضل تقدير.

جدول رقم (5): جذر متوسط مربعات الخطأ (RMSE)
تقدير القيم المفقودة لمعدلات الوفاة - الحالة الأولى

السنة	الدولة	الحالة الأولى - (أ)						الحالة الأولى - (ب)					
		Lin	Pol	PCHIP	CS	DSS	χ^2	Lin	Pol	PCHIP	CS	DSS	χ^2
2020	BEL	0.0023	0.0050	0.0031	0.0034	0.0002	0.2	0.0171	0.0238	0.0103	0.0041	0.0165	0.3
	DEN	0.0005	0.0018	0.0002	0.0002	0.0062	0.1	0.0080	0.0108	0.0080	0.0101	0.0145	0.92
	FIN	0.0004	0.0018	0.0001	0.0005	0.0011	0.19	0.0121	0.0241	0.0056	0.0034	0.0235	0.44
	NOR	0.0004	0.0041	0.0008	0.0013	0.0006	0.11	0.0002	0.0088	0.0002	0.0067	0.0042	0.06
	PRT	0.0039	0.0059	0.0034	0.0031	0.0041	0.35	0.0066	0.0057	0.0106	0.0157	0.0107	19.6*
	SWE	0.0003	0.0006	0.0001	0.0001	0.0001	0.11	0.0045	0.0004	0.0058	0.0130	0.0038	2.6
	SWI	0.0013	0.0006	0.0011	0.0012	0.0008	2.5	0.0234	0.0153	0.0282	0.0312	0.0153	3.5
2019	BEL	0.0004	0.0045	0.0004	0.0002	0.0012	0.09	0.0148	0.0282	0.0130	0.0094	0.0307	1.1
	DEN	0.0018	0.0028	0.0025	0.0032	0.0015	0.31	0.0070	0.0188	0.0070	0.0003	0.0153	0.12
	FIN	0.0010	0.0016	0.0006	0.0005	0.0009	0.10	0.0194	0.0189	0.0196	0.0228	0.0207	4.8
	NOR	0.0004	0.0177	0.0012	0.0011	0.0084	0.31	0.0798	0.1610	0.0792	0.0368	0.0664	2.3
	PRT	0.0021	0.0045	0.0020	0.0020	0.0021	0.20	0.0403	0.0529	0.0382	0.0393	0.0462	1.7
	SWE	0.0017	0.0016	0.0024	0.0028	0.0019	0.07	0.0097	0.0063	0.0171	0.0258	0.0081	3.3
	SWI	0.0004	0.0035	0.0002	0.0008	0.0025	0.07	0.0106	0.0324	0.0106	0.0073	0.0403	0.58
2018	BEL	0.0004	0.0023	0.0010	0.0017	0.0006	0.4	0.0325	0.0302	0.0274	0.0130	0.0378	1.87
	DEN	0.0031	0.0047	0.0031	0.0028	0.0043	0.29	0.0046	0.0033	0.0126	0.0300	0.0012	0.36
	FIN	0.0008	0.0023	0.0016	0.0025	0.0027	0.14	0.0283	0.0491	0.0315	0.1123	0.0097	1.7
	NOR	0.0006	0.0064	0.0004	0.0008	0.0001	0.44	0.1140	0.1056	0.1498	0.1892	0.1379	108.8*
	PRT	0.0011	0.0051	0.0012	0.0011	0.0006	1.33	0.0115	0.0081	0.0217	0.0521	0.0101	0.43
	SWE	0.0036	0.0048	0.0041	0.0047	0.0029	0.43	0.0806	0.0826	0.0784	0.0476	0.0816	25*
	SWI	0.0003	0.0029	0.0001	0.0005	0.0012	0.20	0.0537	0.0376	0.0584	0.0932	0.0454	0.75
		تم اختيار القيم المفقودة لتكون عند الأعمار: 90, 60, 40, 30, 20, 5 * الفروق بين القيم المقدرة والفعلية ذات دلالة إحصائية						تم اختيار القيم المفقودة لتكون عند الأعمار: 104, 41, 29, 22, 18, 4 (2018) - 103, 68, 41, 33, 17, 8 (2019) - 102, 45, 39, 31, 23, 9 (2020)					

من الجدول السابق، يتضح أن أغلب الطرق المستخدمة، باستثناء الاستكمال الخطي، كانت مناسبة لتقدير القيم المفقودة المتقطعة في معدلات الوفاة.

جدول رقم (6): جذر متوسط مربعات الخطأ (RMSE)

تقدير القيم المفقودة لمعدلات الوفاة - الحالة الثانية

السنة	الدولة	الحالة الثانية - (أ)						الحالة الثانية - (ب)					
		Lin	Pol	PCHIP	CS	DSS	χ^2	Lin	Pol	PCHIP	CS	DSS	χ^2
2020	BEL	0.0149	0.0130	0.0239	0.0249	0.0107	1.7	0.0277	0.0299	0.0260	0.0057	0.0199	1.02
	DEN	0.0068	0.0115	0.0078	0.0116	0.0120	0.92	0.0233	0.0251	0.0229	0.0209	0.0148	0.91
	FIN	0.0034	0.0044	0.0008	0.0015	0.0008	0.02	0.0203	0.0352	0.0128	0.0107	0.0343	1.54
	NOR	0.0170	0.0084	0.0266	0.0429	0.0208	0.56	0.0350	0.0161	0.0381	0.0801	0.0444	0.22
	PRT	0.0089	0.0028	0.0162	0.0195	0.0105	5.8	0.0072	0.0107	0.0087	0.0135	0.0163	0.63
	SWE	0.0138	0.0060	0.0204	0.0306	0.0063	0.40	0.0410	0.0293	0.0492	0.0577	0.0381	7.53
	SWI	0.0156	0.0111	0.0231	0.0335	0.0133	5.1	0.0214	0.0185	0.0229	0.0253	0.0191	2.88
2019	BEL	0.0035	0.0154	0.0021	0.0041	0.0122	0.15	0.0599	0.0557	0.0554	0.0679	0.0577	38.9*
	DEN	0.0100	0.0068	0.0220	0.0294	0.0113	5.7	0.0791	0.0775	0.0830	0.0496	0.0687	27.4*
	FIN	0.0096	0.0115	0.0092	0.0088	0.0103	1.5	0.0242	0.0215	0.0305	0.0534	0.0287	5.7
	NOR	0.0308	0.0473	0.0406	0.0497	0.0356	13.2	0.2792	0.2831	0.1919	0.1864	0.2685	304*
	PRT	0.0186	0.0138	0.0246	0.0367	0.0253	8.7	0.0674	0.0659	0.0629	0.1260	0.0742	41.1*
	SWE	0.0079	0.0105	0.0081	0.0170	0.0088	10.1	0.0091	0.0134	0.0332	0.1184	0.0131	1.1
	SWI	0.0103	0.0178	0.0091	0.0079	0.0172	0.03	0.1457	0.1276	0.1458	0.2248	0.1463	209.7*
2018	BEL	0.0116	0.0021	0.0194	0.0261	0.0308	4.96	0.0389	0.0322	0.0422	0.1006	0.0472	11.5
	DEN	0.0042	0.0037	0.0056	0.0081	0.0071	4.41	0.0199	0.0171	0.0223	0.0166	0.0253	3.08
	FIN	0.0331	0.0311	0.0397	0.0479	0.0313	47.5	0.0448	0.0322	0.0619	0.1758	0.0397	5.03
	NOR	0.0035	0.0259	0.0108	0.0255	0.0108	0.22	0.1043	0.1134	0.1067	0.0971	0.0994	84.5*
	PRT	0.0024	0.0120	0.0053	0.0136	0.0025	0.11	0.0311	0.0308	0.0322	0.0658	0.0328	6.1
	SWE	0.0066	0.0086	0.0118	0.0175	0.0118	0.75	0.0831	0.0804	0.0846	0.2005	0.0782	59.1*
	SWI	0.0076	0.0237	0.0051	0.0097	0.0195	0.46	0.0506	0.0426	0.0518	0.0849	0.0453	15.8*
		تم اختيار القيم المفقودة لتكون عند الأعمار: 99, 98, 51, 50, 13, 12 * الفروق بين القيم المقدرة والفعلية ذات دلالة إحصائية						تم اختيار القيم المفقودة لتكون عند الأعمار: 104, 103, 30, 29, 19, 18 (2018) - 104, 103, 34, 33, 22, 21 (2019) - 102, 101, 23, 22, 10, 9 (2020)					

ومن النتائج السابقة، يتضح أنه عند تقدير أو استكمال القيم المفقودة في بيانات معدلات الوفاة، لم تكن هناك أفضلية واضحة لأي من الطرق المستخدمة، كما كانت التقديرات جميعها في حالة القيم المفقودة التي لا تقع في مناطق توجد عندها تقلبات ملحوظة في منحنى البيانات (سواء في الحالة الأولى أو الثانية) مقبولة. كما يمكن ملاحظة أن الاستكمال الخطي لم يكن مناسبا في حالة القيم المفقودة التي تقع في مناطق توجد عندها تقلبات ملحوظة في منحنى البيانات (سواء في الحالة الأولى أو الثانية)، وكانت المفاضلة بين باقي الطرق فقط.

وبالتالي، يتضح مما سبق أن جميع الطرق التي تم استخدامها في هذه الدراسة، لتقدير أو استكمال القيم المفقودة سواء في أعداد الوفيات أو معدلات الوفاة، تصلح بشكل أو آخر لهذا الغرض، مع ملاحظة ما يلي:

1. أن زيادة دقة التقدير تعني بالضرورة انخفاض درجة النعومة أو التمهيد (smoothness)، والعكس صحيح، وبالتالي إذا كان أحد أهداف التحليل هو تمهيد بيانات الوفيات، فإنه يفضل الاعتماد على الطرق التي تعطي تقديرات أكثر نعومة كالشرائح الجزائية المقطعة (DSS) واستكمال Hermite (PCHIP).
2. أن طريقة شرائح التمهيد المقطعة (DSS)، والتي تعتمد في الأصل على تمهيد القيم، تتميز بقدر أعلى من المرونة مقارنة بالاستكمال الخطي أو الاستكمال بكثيرات الحدود التكعيبية، حيث يمكن التحكم في مقدار التمهيد (وبالتالي دقة التقدير) من خلال تغيير قيم معلمة أو أكثر في النموذج. فمثلا، عند تغيير قيمة معلمة التمهيد λ المحسوبة على أساس معيار التحقق المتقاطع المعمم (GCV)، تصبح تقديرات القيم المفقودة أكثر دقة، مع كونها ممهدة في نفس الوقت. ويعرض الجدول رقم (7) قيم RMSE عند استخدام قيمة مختلفة لمعلمة التمهيد (λ) في إحدى حالات تقدير القيم المفقودة في معدلات الوفاة، على سبيل المثال.

جدول رقم (٧): جذر متوسط مربعات الخطأ عند تقدير القيم المفقودة لمعدلات الوفاة الحالة الأولى (ب) مع تعديل قيمة λ في طريقة DSS

السنة	الدولة	Lin	Pol	PCHIP	CS	DSS λ_{gcv}	DSS $\lambda = \frac{\lambda_{gcv}}{500}$
2019	BEL	0.0148	0.0282	0.0130	0.0094	0.0307	0.00880
	DEN	0.0070	0.0188	0.0070	0.0003	0.0153	0.00107
	FIN	0.0194	0.0189	0.0196	0.0228	0.0207	0.01525
	NOR	0.0798	0.1610	0.0792	0.0368	0.0664	0.04142
	PRT	0.0403	0.0529	0.0382	0.0393	0.0462	0.04283
	SWE	0.0097	0.0063	0.0171	0.0258	0.0081	0.00528
	SWI	0.0106	0.0324	0.0106	0.0073	0.0403	0.00217
		القيم المفقودة عند الأعمار: 103, 68, 41, 33, 17, 8					

3. أن تقديرات القيم المفقودة إذا كانت واقعة عند حدود البيانات (data boundaries) باستخدام طريقة شرائح التمهيد المقطعة قد لا تكون مناسبة في حالة العينات الصغيرة إذا كانت معلمة التمهيد محسوبة على أساس معيار التحقق المتقاطع المعمم، حيث قد يقتضي الأمر تعديل قيمة تلك المعلمة بما يتناسب مع طبيعة البيانات (Garcia, 2010).

4. أن الاستكمال بكثيرات الحدود التكعيبية، لم يكن مناسبا في أغلب الحالات، وذلك لأن المعلمات يتم تقديرها باستخدام جميع قيم البيانات والتي تتفاوت كثيرا حسب العمر وفقا لطبيعة معدلات الوفاة، حيث كانت التقديرات عند أعمار محددة إما بعيدة جدا عن القيم الأصلية أو غير موجبة.

5. أن الاستكمال الخطي، وبغض النظر عن التمهيد، يعطي تقديرات مقبولة للقيم المفقودة إذا كان تباين القيم الأصلية ليس كبيرا أو كانت التقلبات في منحنى البيانات محدودة، وكذلك إذا كانت المسافات بين القيم المفقودة والقيم المعلومة الموجودة حولها في نطاق البيانات ليست بعيدة.

7. التوصيات:

1. دراسة مدى ملاءمة تطبيق بعض الطرق التي تعتمد على التحليل الرياضي (زايد، م.، الأشقر، ا.، 2020) أو بعض النماذج التي يتم فيها إيجاد القيم المقدرة بطريقة تعطي وزنا أكبر لنقاط البيانات القريبة من القيم المفقودة، مثل الانحدار المحلي (local regression)، والمقارنة بينها وبين الطرق المستخدمة في هذا البحث.

2. المقارنة بين الأساليب المستخدمة في هذا البحث، وغيرها، لتقدير القيم المفقودة في بيانات الوفيات عند الأعمار الصغيرة والكبيرة، وخاصة عند حدود البيانات، وكذلك في حالة السلاسل الزمنية المقطعية.

3. العمل على إنشاء قاعدة بيانات خاصة بالوفيات والسكان في الدول العربية، وجعلها متاحة لأغراض البحث العلمي وتحديثها بشكل مستمر.

المراجع:

زايد، محمد عبد اللطيف، الأشقر، السيد الشربيني (2020). المدخل التباينى فى التحليل الرياضى كطريقة حديثة لتسوية معدلات الوفاة. المجلة العلمية للدراسات والبحوث المالية والتجارية، 1(العدد الثانى - الجزء الثانى)، ٥٤٩-572.

الأشقر، السيد الشربيني، زايد، محمد عبد اللطيف (2020). دراسة مقارنة لثلاث طرق لاستكمال جداول الحياة المختصرة. مجلة البحوث المالية والتجارية، جامعة بورسعيد، 21(4-1)، 474-494.

محمد حبيب، ا.، حافظ محمد، م. (2011). مقارنة بعض طرائق تمهيد الانحدار اللامعلمي باستخدام المحاكاة. مجلة القادسية لعلوم الحاسوب والرياضيات، 3(2)، 1-19.

Acal, C., Escabias, M., Aguilera, A. M., & Valderrama, M. J. (2021). COVID-19 Data Imputation by Multiple Function-on-Function Principal Component Regression. Mathematics, 9(11), 1237.

Andreopoulos, P., Bersimis, G. F., Tragaki, A., & Rovolis, A. (2019). Mortality modeling using probability distributions. Application in Greek mortality data. Communications in Statistics-Theory and Methods, 48(1), 127-140.

Azizan, I., Karim, S. A. B. A., & Raju, S. S. K. (2018). Fitting rainfall data by using cubic spline interpolation. In MATEC Web of Conferences (Vol. 225, p. 05001). EDP Sciences.

Bazo-Alvarez, J. C., Morris, T. P., Pham, T. M., Carpenter, J. R., & Petersen, I. (2020). Handling Missing Values in Interrupted Time Series Analysis of Longitudinal Individual-Level Data. Clinical Epidemiology, 12, 1045.

Currie, I. D., Durban, M., & Eilers, P. H. (2004). Smoothing and forecasting mortality rates. Statistical modelling, 4(4), 279-298.

Garcia D. (2010). Robust smoothing of gridded data in one and higher dimensions with missing values. Comput Stat Data Anal, 54(4), 1167-1178.

Human Mortality Database (2021). University of California, Berkeley (USA), and Max Planck Institute for Demographic Research (Germany). (www.mortality.org)

McNeil, N., Odton, P., & Ueranantasun, A. (2011). Spline interpolation of demographic data revisited. Sonklanakarin Journal of Science and Technology, 33(1), 117.

Meseguer, A. (2020). Fundamentals of Numerical Mathematics for Physicists and Engineers. John Wiley & Sons.

Rabbath, C. A., & Corriveau, D. (2019). A comparison of piecewise cubic Hermite interpolating polynomials, cubic splines and piecewise linear functions for the approximation of projectile aerodynamics. Defence Technology, 15(5), 741-757.

Schmertmann, C. (2021). D-splines: Estimating rate schedules using high-dimensional splines with empirical demographic penalties. Demographic Research, 44(45), 1085-1114

Siauw, T., & Bayen, A. (2015). An introduction to MATLAB® programming and numerical methods for engineers. Academic Press.

Simos, T. E., Tsitouras, C., Kovalnogov, V. N., Fedorov, R. V., & Generalov, D. A. (2021). Real-Time Estimation of R0 for COVID-19 Spread. Mathematics, 9(6), 664.

Zaghiyan, M. R., Eslamian, S., Gohari, A., & Ebrahimi, M. S. (2021). Temporal correction of irregular observed intervals of groundwater level series using interpolation techniques. Theoretical and Applied Climatology, 1-11.

Wahba, G. (1990). Estimating the smoothing parameter: Spline models for observational data. Society for Industrial Mathematics, Philadelphia, 45-65.

Estimation of Missing Values with Application to Mortality Data - A Comparative Study

Mohammad Zayed

Assistant Professor, Quantitative Methods Department,
School of Business, King Faisal University
Lecturer, Applied Statistics and Insurance Department
Faculty of Commerce, Mansoura University
m.a.zayed@mans.edu.eg

Abstract:

Estimating premiums in personal insurance as well as demographic planning requires accurate and complete mortality data at different ages. In some cases, there may be missing values in the data, which makes estimating or interpolating those values an important issue in actuarial sciences and demography and of interest to specialized researchers. This study compares several mathematical and statistical interpolation methods, such as linear interpolation, piecewise cubic Hermite interpolating polynomial (PCHIP), and discretized smoothing splines, to interpolate missing mortality data, whether in deaths or death rates. The methods were applied to mortality data for seven European countries for the period 2018-2020 assuming different cases of missing values. In most cases, except for polynomial interpolation, results were acceptable, bearing in mind that some methods may be more appropriate in the case of homogeneous data when data points are close. It is also preferable to use methods that are based primarily on smoothing if the goal is to obtain smooth values. It is recommended to compare the methods applied in this research and others, to estimate missing values in mortality data for young and old ages, especially at data boundaries, and for cross-sectional time series.

Keywords: Linear interpolation - Piecewise cubic Hermite interpolating polynomial (PCHIP) - cubic splines - discretized smoothing splines.