

TABLE OF CONTENTS

Acknowledgements	ii
Editor's Introduction	iii
Contents	v
List of Tables	vi
1 Introduction	1
2 The City of Viana do Castelo	2
3 The Manuscript Sources	10
4 Record Linkage	17
5 Preparation of the Viana Data	25
6 Record Linkage of the Viana Data	31
7 Summary	56
Appendix A Facsimiles of Sources	58
Appendix B Data Recorded on the Electoral Registers	62
Manuscript Sources	64
Printed Sources	65
References	65

LIST OF TABLES

1	The Population of Viana in the Nineteenth Century	9
2	Portuguese 19th Century Electoral Legislation	14
3	Example Component Name Standardisation	31
4	The Life History of João da Silva São Miguel	34
5	Full Name Repetition on the Cemetery Lists	36
6	Match Scoring Functions	41
7	Stage 1 Record Linkage Statistics (Males)	44
8	Stage 1 Record Linkage Statistics (Females)	44
9	The Effect of Ignoring Component Name Order for all but First Component Names (Males)	44
10	Stages 2, 3, & 4 Record Linkage Statistics	47
11	Final Record Linkage Statistics	47
12	Illustrative Example Linkage 1: Ordering of Blocks	50
13	Illustrative Example Linkage 1 Division of Blocks into Sub-Blocks	51
14a	Illustrative Example Linkage 1: Extracted Historical Persons	52
14b	Illustrative Example Linkage 1: Records Returned to Main Chain	53
15a	Illustrative Example Linkage 2. Division of Blocks into Sub-Blocks	54
15a	Illustrative Example Linkage 2 Division of Blocks into Sub-Blocks	55

1. INTRODUCTION

The paucity of aggregate demographic statistics for nineteenth century Portugal results in the need to exploit fully various sources of supplementary data. An increasingly popular approach involves the reconstitution of one town or city using a database of several manuscript sources. This enables micro-analyses to be performed which enhance the understanding of secular demographic trends derived from those aggregate statistics that are available.¹ However, these studies raise major caveats in their representativeness (Schofield, 1972, Levine 1976, Åkerman, 1977). Also, they provide a picture of just one locality, and no regional patterns can be identified, they are held to be non-comparable and non-cumulative, so that any amount of such micro-studies would not enable the reconstruction of the macro-structure of a whole society (Macfarlane, 1977). Nevertheless, within these constraints, they can be extremely informative about a particular society.

The Port-city of Viana do Castelo provides an ideal opportunity for the investigation of secular demographic trends through the reconstitution of its population from the abundance of records generated by the Portuguese administrative system. With a population of just less than 10,000 during the nineteenth century, it is large enough to limit the effects of random variation, while not too large to preclude a close examination of the individual level data. Also, its role as the administrative centre and principal port of the District of Viana offers the study of stage migration of the rural peasantry, the migrations of the urban poor, and the migrations of the urban elite.

This paper, drawn from the research of Reis (1987), Doulton & Kitts (1988), and Kitts (1988), describes the reconstitution of Viana; initially, this reconstitution is restricted to the male urban elite, for whom record linkage is most accurate. The paper is structured as follows. Section 2 covers the economic and demographic history of Viana, and focuses on local and national events in the nineteenth century, in order to establish the circumstances under which the sources being used were created. In Section 3 the manuscript sources (muster-rolls, electoral registers, passport books, and cemetery lists) on which the reconstitution is currently based are described in detail. Section 4 introduces reconstitution methodology with a brief review of record linkage techniques and studies, and a description of recently-developed software which greatly facilitates the record linkage process. In Section 5, the preparation of the Viana data for record linkage is described, this Section covers the problems arising

1 Examples are the reconstitution of a selection of English parishes (see Wrigley & Schofield, 1973), a selection of Swedish parishes (see Danell, 1981, or Sundin, 1984), the province of Quebec (see Gauvreau, 1986), Eindhoven, Netherlands (see Dijk, 1977)

in the use of the manuscript sources, and the standardisation and coding techniques used to overcome these problems. In Section 6, the record linkage of the Viana data is described in detail. Finally, in Section 7, the methodological and substantive issues raised in this paper are summarised.

2. THE CITY OF VIANA DO CASTELO

The City of Viana is situated on the northern bank of the mouth of the River Lima which runs westwards through Minho, the northernmost province of Portugal. The City, whose urban centre comprises the two parishes of Santa Maria Maior da Matriz, and Nossa Senhora de Monserrate, serves as the administrative centre of both the Borough and the District of Viana. Historically, Santa Maria Maior, the larger of the two parishes, housed the prosperous trading community of the Town, while fishing activities were centred in Monserrate.

2.1 The History of Viana

The ancient Borough of Viana da Foz do Lima² (literally Viana at the mouth of the Lima) was founded by Dom Afonso III on 18 June 1258 (confirmed in 1262), at which time the settlement itself was elevated to the category of *vila* (town) (Moreira, 1984). By the end of the same century both the church and town walls (surrounding what was later to become the commercial centre of Santa Maria Maior) had been completed (Guerra, 1880).

By the end of the fourteenth century Viana had not grown significantly. Frei Martinho do Amor de Deus, referring to the population living along the bank of the River at that time, wrote that 'Viana was a poor town composed of small houses which would more appropriately be described as humble cottages'.³ José Caldas (1919) elaborates by describing Viana as consisting of only a small community of fishermen and sailors.

The Town's development began with the period of discoveries in the late fifteenth century when economic activities became concentrated around the coastal areas of Portugal. The earliest trade links between Viana and the ports of northern Europe are reputed to have been founded by a Jewish community from Aragon who settled in the Praça Velha⁴ – the centre of the Town (Caldas, 1919). Initially, international trade was served by local ships, but over time, foreign vessels became more involved.

² This original Borough covered an area of about 180 km² bounded by the rivers Âncora (north) and Lima (south). The parishes to the south of the Lima were incorporated in the Borough of Viana with the administrative reforms of 1835.

³ Translated from Abel Viana (1953: 7).

⁴ Spanish Jews were being persecuted in the late fifteenth century and it was not uncommon for them to seek refuge in Portugal (Livermore, 1966).

Abel Viana (1953) describes Viana during the reign of Manuel I (1495–1521) as having around seventy ships of various types ⁵, which gave the Town a liveliness comparable with only the most important of European port cities. Writing at that time, Frei Luís de Sousa noted that the nobles of Viana, like those of Venice and Genoa, were engaged in trade and shipping activities, in contrast with most of their contemporaries throughout the rest of Portugal,⁶ in the same vein, Flavio Gonçalves writes of the aristocratic seaport of Viana.⁷ Trade in silk and porcelain from the Far East, glass from Venice, and fish, sugar, and timber from Brazil (particularly sugar from Pernambuco) enriched the Town, and it soon expanded beyond its walls.⁸ Other import-export goods included coal, iron, lime, cotton, leather, cod, rice, salt, wheat, and wine. The growing importance of the Town also led to the development of its watch tower into a small castle.⁹

During the sixteenth century, the importance of Viana as a sea port rested on its position at the centre of an important net of commercial activities including long distance trade links with the Far East, Africa, Brazil, and most of Northern Europe (including Russia), together with other Iberian ports such as Aveiro, Figueira da Foz, Lisboa, Porto, Setúbal, and those of Spanish Galicia. The international links, besides providing extensive trading opportunities, were also responsible for the introduction of new crops and industries to the District of Viana.¹⁰ It was at this time that the commercial centre of Viana shifted from the Praça Velha to the Campo do Forno, but smaller merchant businesses remained in the Ruas do Caes and Postigo, and in the Porta Principal.

There is little evidence to suggest that the Spanish domination of Portugal (1580–1640) affected activities in Viana, but the effects of the Wars of Restauração (1640), coupled with increasingly fierce competition from Dutch and English ships for the control of international trade, were responsible for the crisis which first appeared in the mid-seventeenth century.

By the 1680s, however, despite the reduction in importance of the sugar trade with Brazil, signs of recuperation accompanied the beginning of the gold rush in Brazil and the increase in wine and cattle exports to Northern Europe (Moreira, 1984).¹¹ With the end of the wars in 1713, Viana's situation improved along with that of the country as a whole, trade flourished again, particularly with Brazil, and another era of prosperity commenced (Crespo, 1957: 17).

The Town continued to enjoy this renewed prosperity until the middle of the

5 There is, however, some disagreement over the number of vessels engaged in specific activities at this time (see Castro, 1979: 21).

6 Cited by Abel Viana (1953: 9).

7 See Sayers (1968: 455, cited by Pierson, 1970: 459).

8 The Ruas da Bandeira, Picota, Carreira, and São Sebastião, and the Bairros da Ribeira, São João de Arga, Abelheira, etc. date from this period (Crespo, 1957).

9 Ordered by the Decree of 1548.

10 Links with Ireland and the Low Countries, for example, resulted in the introduction of the lace industry (Viana, 1953).

11 The Treaty of Methuen (1703), binding Portugal to admit British cloths on the same footing as before prohibition, and Great Britain to admit Portuguese wines at two-thirds the duty paid for French wines, was partly responsible for this increase in exports. The Treaty is described by Smith (1776).

eighteenth century Braga (1985) reports that between 1720 and 1741, 1018 ships entered the Port of Viana, of which 10% were Portuguese, 43% English, and 22% French. Of these, 29% were carrying cod from Newfoundland, 26% carried iron, and 12% lime. In comparison, between 1749 and 1774, only 855 ships entered the port, of which 9% were Portuguese, 40% English, and 35% Galician. Of these, 28% carried cod, 26% iron, 6% wheat and less than 2% lime. Besides suggesting a decrease in the volume of trade, these figures clearly demonstrate the dominance of foreign vessels by the eighteenth century, and show large variations in the quantities of particular goods traded. In 1770, Viana imported cattle from Galicia, olives and wheat from Greece, iron from Switzerland, hardware from England, canvas and linens from Russia, and tea from India. In the same year, its principal exports were corn, oranges, rye, and wine (Loureiro, 1923: 25).¹²

Towards the end of the eighteenth century, the pressure of competition from Porto proceeded to erode Viana's status and wealth still further, despite attempts to revitalise the economy by the local bourgeoisie. At the same time, the silting of the bar began to deny the larger more modern ships access to the port (Crespo, 1957). Also, during the early 1770s, as a result of an investigation into contraband trade, the Marquis de Pombal rescinded the Port's licence to conduct international trade.

The demise of Viana continued into the nineteenth century. Local commerce was ruined by the disruptions of war that followed the French invasions of 1807, 1809, and 1810, and Viana's role as an entrepot port subsequently declined, especially after the independence of Brazil (1822).

Nevertheless, while 'there is no doubt that the moment of glory of the Port of Viana was during the period of discoveries' (Sampaio, 1981), Viana entered the nineteenth century as one of the largest towns in the country, and was even elevated to the status of cidade (city) in 1848, when its name was changed to Viana do Castelo.¹³ However, before examining nineteenth century Viana more closely, it is necessary to discuss Portugal as a whole, in order to provide a global picture from which to focus.

2.2 Portugal in the Nineteenth Century

The upheavals of the French Revolution and Napoleonic conquests affected the pattern of Europe's development deeply. With respect to Portugal, it is unlikely that Trend (1957: 175) was exaggerating when he wrote that her recovery from the Napoleonic War took more than a century to accomplish, and an outline of some of the most important events in Portugal therefore forms an essential part of this study, establishing the conjuncture under which the sources described in Section 3 were originally created.

¹² Cited by Sampaio (1981: 14).

¹³ The change in name was associated with the role played by the Town's small castle during the Patuleia War (1846–1847) (Feijó, 1983). Some confusion is unavoidable here – where Viana is referred to in periods before, spanning, or following the year 1848, it will be referred to as the Town, Town/City, and City of Viana, respectively.

First, the flight of Portuguese government to Rio de Janeiro and the extended residence of the monarchy there until 1820 – a direct cause of the French invasions of 1807, 1809, and 1810 – undoubtedly accelerated the inevitability of Brazilian independence which was to deprive Portugal of the fabulous profits previously enjoyed from the gold and diamonds mined there¹⁴ Second, a less direct effect of the Napoleonic Empire was the diffusion of liberal ideas which, often proceeding in the form of anti-British sentiment (Silva, 1985 397), culminated in the Revolution of 1820 By the end of the same year, the first version of the constitution had been proclaimed, and Portugal had seen her first parliamentary elections

The Revolution of 1820 also forced the return of Dom João VI from Brazil in 1821 He left his eldest son Pedro as Regent in Brazil, and brought with him his second son, Miguel

In the following year, Prince Pedro proclaimed Brazil independent,¹⁵ and became its first Emperor This had an immediate impact on the Portuguese economy, for example, the instant drop in exports of wine to Brazilian markets stirred up unrest in the provinces of Minho and Trás-os-Montes (Ruiz, 1980) Such unrest ensured that the constitution elaborated by the elected cortes in 1822 did not survive long

In May 1823, Miguel led a military revolt, later setting up a new government with his father The cortes was dissolved, and the bulk of the legislation passed by the Liberals was repealed Further, Miguel, striving to establish a regime 'untainted by constitutionalism' (Nowell, 1952 185), led an abortive *coup d'état* in 1824, this resulted in his exile

Upon the death of João VI in 1826, his son Pedro – the first emperor of Brazil, abdicated the Portuguese throne in favour of his seven year old daughter, Maria da Gloria, and decided that she should marry his younger brother Miguel He decreed and sent to Portugal a new constitution, the Charter, which Miguel was to accept and swear to maintain The Charter was, however, very coldly received in Lisboa and the interior, and Miguel, having solemnly partaken in a ceremony of betrothal to Maria da Gloria and sworn by the Charter in Vienna, and having been greeted with popular enthusiasm upon his return to Lisboa in 1828, seized the throne of Portugal¹⁶ Miguel's absolutist measures swiftly provoked military reaction in Porto, and unrest soon spread to most of the cities north of the Mondego River¹⁷ plunging the country into prolonged civil war (Saraiva, 1978 266)

The ensuing tyranny of Dom Miguel is recorded as having been singularly bloody, scarred by innumerable arrests, deportations, and executions,¹⁸ and hav-

14 However, the political independence of Brazil by no means produced economic independence, since Portuguese merchants and traders remained important in Brazilian life for a long time, controlling much of the trade (Serrão, 1974 51–52)

15 Brazil was declared independent on 7 September 1822

16 While Caldas (1919 693) states that Dom Miguel was officially unrecognised by the Pope in 1829, Nowell (1952 187) writes that he was given a sort of recognition by many governments, including Russia, France, Spain, the United States, and the papacy

17 The Mondego River flows west through the City of Coimbra

18 Between 1828 and 1833, 618 political prisoners were jailed (Saraiva, 1978 266), and

ing instigated a considerable flow of emigration, both politically motivated and otherwise (Oliveira Martins, 1891)

In 1831 Pedro abdicated his Brazilian throne, and set out for the Açores with the intention of restoring his daughter to the Portuguese throne. Less than two years later, in 1833, the Liberals landed near Porto, and in July of the same year they entered Lisboa, meeting little resistance. Miguel was eventually forced to make peace almost a year later, in May 1834, and was banished by Pedro, who enthroned his daughter Maria before dying later that year.

Thus, a constitutional government, formed around a core of moderate liberals, was finally re-established in 1834. There followed the first phase of liberalism during which the reformist Government focused on political and institutional matters, rather than the economic problems that the country faced (Feijó, 1983: 2).

In 1836, the radical liberal Septembristas (as opposed to the Chartistas, who were in favour of Pedro's constitutional Charter), were elected to the north, and a *coup* shortly thereafter brought them to power (without deposing Queen Maria). A subsequent revolt, in 1837, led by the Chartistas' Dukes of Terceira and Saldanha, was suppressed, but the seesaw situation did result in the compromise constitution of 1838 – the third constitution of Portuguese liberalism.

By 1842, under António Bernardo da Costa Cabral, the Chartistas had regained power, and restored the Charter of 1826 once more. But, unpopular progressive reforms imposed by the government of Costa Cabral led to the rising of Maria da Fonte in 1846,¹⁹ which spread from the Minho throughout Portugal, bringing the Government's replacement by Septembristas once more.

Tension remained high in the summer of 1846, and broke in October when a *coup* led by the Duke of Saldanha initiated another civil war – the Patuleia War – which rekindled the bitter rivalries of the 1828–1834 war. Dom Miguel was acclaimed once more. This civil war continued until June 1847, when Spanish intervention accelerated the end of hostilities, resulting in a further period of political instability, and the return of Costa Cabral.

Only after 1851, when the Government of Costa Cabral fell again, did a stable form of parliamentary rule finally emerge in Portugal, with the Regeneration Party²⁰ (previously the Chartistas, excluding the followers of Costa Cabral) initially in power, and the Progressive Party (previously the Septembristas) in opposition.

In contrast, the second half of the nineteenth century has been described as disappointingly inert, with the liberals evolving into a conservative group, and many of the promises of the Liberal Revolution left unfulfilled.

The third quarter of the nineteenth century saw the second phase of liberalism,

115 executions are documented in detail (Oliveira Martins, 1891: 179–180).

¹⁹ Recent authors (e.g. Brandão & Rowland, 1980), however, place the roots of the discontent which led to the peasantry's sudden mobilisation as a 'headless body, all life and soul, all popular will' (Roby, 1846: 20, cited by Villaverde Cabral, 1976), more deeply in the division of property, the key to many of the tensions that existed in the Minho.

²⁰ The Regeneração has been popularised as the 'Portuguese name for Capitalism' after Oliveira Martins (Feijó, 1983: 41).

when some attempts were made to improve the economy, principally through the construction of modern communication and transport networks (Feijó, 1983 2) The period 1851–1865 saw the Regeneration and Progressive Parties in power for six and eight years respectively, after which a coalition cabinet was formed. However, although several bodies of legislation concentrating on the provision of an economic infrastructure were passed between 1851 and 1870,²¹ by the latter date liberalism was being accused of being incapable of replacing Old Portugal by a new form of life (Feijó, 1983 43)

During the 1890s, political and economic disturbances in Brazil, associated with the proclamation of the Brazilian Republic and the fall in coffee prices, caused a dramatic fall in the exchange. This had a disastrous effect on the value of the remittances from Portuguese emigrants, causing difficulties in obtaining fresh advances for Portugal, reducing the means with which the interest on foreign debts was paid, and leading to the outbreak of unrest in Porto in January 1891. The crisis was aggravated by the British Ultimatum concerning Portuguese interests in Africa (Trend, 1957 181)

In conclusion, nineteenth century Portugal has presented historians with the paradox that amidst some major developments, the Country's economic conditions continued to deteriorate. The most popular reasons put forward for this include the far-reaching effects of the French invasions, the fragile economy that was so heavily dependent on foreign countries,²² the limited effectiveness of the land reforms on agriculture, and, the rigidity and inertia of Portuguese social structure (Mendes, 1980, Reis, 1984)

2.3 Viana in the Nineteenth Century

The drop in exports of wine to Brazilian markets in particular augmented the social tension in the Minho and Trás-os-Montes and is deemed to be a factor of great importance in explaining the Miguelista reaction of northern Portugal. At the same time the drop in Brazilian demand for manufactured products (cotton, silk, wool, and iron utensils) increased the fragility of the already retarded development of industry. Furthermore, the situation was made worse by the impotence of the bourgeois in bringing about political and economic reform (Ruiz, 1980 795)

It is not surprising then that the commercial Town of Viana received Pedro's Charter of 1826 as coldly as anywhere, thereby causing an increase in the popularity of Miguel. His return to Portugal in 1828 was greeted with enthusiasm in the Town, as was his visit on 30 March 1832 (Caldas, 1919)

Three days after troops loyal to the queen entered the Town in March 1834, the Miguelite administration of Viana was replaced by a temporary interim committee (Feijó, 1983 299)

The radical position of the Town was however reaffirmed in 1846, when the Government headed by Saldanha (replacing the dismissed Government of

21 For example, the property reforms of the early 1860s, and particularly the Civil Code of 1867

22 Great Britain absorbed 50–60% of Portugal's external produce (Halpern Pereira, 1971)

Palmella which had been appointed during the rising of Maria da Fonte) reinstated many followers of Costa Cabral. At the outbreak of the Patuleia War, Viana adhered to the rebel *Junta* of Porto, relying on the word of honour of the officers of the local garrison that there was no fear of a military revolt in favour of Saldanha. Nevertheless, rebellion within the garrison did break out on 20 October 1846 (the rebellion of Pinhotes), and military officers took over the administration of the Town (which had fled). Within two days, however, almost 3,000 peasants had surrounded the castle, and the rebels were left with little choice but to negotiate a surrender (Feijó, 1983 323–326).

It is a little ironic then that the inhabitants of the Town were to enjoy the new title of 'the City of Viana do Castelo' after 1848, which was awarded because of the valiant defence of the castle by troops loyal to Dona Maria II, defending against Miguelistas!

Returning to Viana's economy, as described in Section 2.1, her importance began to decline in the first half of the nineteenth century, a period during which population stagnation could be identified throughout the Minho. Feijó (1983 54) writes that, in contrast with a rapid national growth, the Borough of Viana witnessed a decelerating growth until the 1880s, caused by the failure to industrialise, political and military occupation by foreign powers, the continuous fighting that followed the liberal revolution of 1820, and the loss of the Brazilian trade.

At the same time, the development of technology and good communication and transport systems was notoriously slow. The first (wooden) bridge over the Lima was only completed in 1819, twelve years after its construction was initially approved, this was only replaced by an iron bridge (designed by Eiffel, and still standing today) in 1878 when the railway from Porto reached the City (Filgueiras, 1979). There was extensive resistance to the adoption of the metric system in 1852 (Feijó, 1983 306). Even the harbour at Viana, which had hitherto played such an integral role in the economy of the entire District, suffered from insufficient funds for its profitable development,²³ resulting in fears that the trade that remained would be driven elsewhere by increases in the bar-tax. A final illustration is that the streets of Viana were first illuminated using gas lanterns only in 1883, more than 35 years after they had been introduced in Lisboa, by the same year, Lisboa had already benefited from electric lighting for five years.²⁴

In conclusion, the economy of Viana in the nineteenth century reflected the transition from bustling international port to regional administrative centre. Viana gradually changed from a flourishing town, '*uma nova Lisboa*' (a new Lisboa) according to Frei Luís de Sousa,²⁵ part of one of Portugal's most prosperous regions, to a peripheral and backward town, incapable of competing with Porto as a port city.

23 *A Aurora do Lima* (12 May 1880)

24 *Viana. A Aurora do Lima* (1 April 1883), Lisboa. Serrão (1970)

25 Quoted by Crespo (1957 16)

2.4 The Population of Viana

This Section describes the overall trends in the population of Viana to the end of the nineteenth century

The enumeration of 1527 refers to 962 households and 3,800 inhabitants of the Town and its surrounding area, a relatively high figure for the time, with 679 households of fishermen and sailors in Monserrate, compared with 146 in Santa Maria Maior (Crespo, 1957)

Moreira (1984) identifies three distinct periods of demographic change during the sixteenth and seventeenth centuries 1517–1580 and 1580–1640, during which the population increased by 86.5% and 109% respectively, and 1640–1707, the period coinciding with the first economic crisis in Viana, during which the population decreased by 31%. Thus, while there is no record of plagues or wars having had any great demographic impact, the effects of economic crises on mortality and out-migration are likely to have been important

In fact, Viana's commercial role was both the cause and effect of its demographic growth, and so, with its decline during the late eighteenth and early nineteenth centuries, out-migration assumed greater importance than ever before

The populations of Viana recorded at the censuses taken during the nineteenth century are shown in Table 1, the figures show that while the population of Santa Maria Maior increased at each point, the population of Monserrate fluctuated considerably

Table 1. The Population of Viana in the Nineteenth Century

Year	Santa Maria Maior			Monserrate		
	Population (1)	Households (2)	(1)/(2)	Population (1)	Households (2)	(1)/(2)
1801	4,825	1,075	4.49	2,444	661	3.70
1864	5,333	1,185	4.50	3,930	868	4.52
1878	5,450	1,211	4.50	3,366	846	3.98
1890	5,590	1,253	4.46	4,092	979	4.18
1900	5,625	1,222	4.60	4,467	917	4.87

Since part of the growth between 1801 and 1864 can be attributed to the large under-registration of the 1801 census,²⁶ it can be seen that Viana experienced very little demographic growth during the nineteenth century, barely achieving a population of 10,000 by 1900. This also suggests that there was little urbanisation in the District of Viana, as the natural increase in the population of the surrounding area must have been mostly absorbed by out-migration. Finally, the figures demonstrate that Monserrate grew far more than Santa Maria Maior during the period. Unfortunately however, no information on the availability of housing and accommodation is available for an analysis of the details of this growth

²⁶ Described in detail by Reis (1987:98)

Considering migration to and from Viana, it is not possible to draw a detailed picture because so very little is known about the various components of these flows. In-migration is known to have been substantial, since almost 30% of the population of the City were non-natives of the Borough in 1890. But, out-migration, both to other parts of Portugal and emigration from Portugal, have hitherto remained unquantified. Nevertheless, it is suggested (Kitts, 1988) that out-migration played a considerable role in the population stagnation of nineteenth century Viana identified both throughout the surrounding area by Feijó (1983), and in the City itself by Reis (1987). The high levels of emigration which are known to have operated throughout north-western Portugal during the second half of the nineteenth century are likely to have accounted for a substantial amount of out-migration, out-migration to other parts of Portugal, particularly Braga and Porto, the two nearest most important urban centres, is also likely to have been important.

3. THE MANUSCRIPT SOURCES

This Section will describe the manuscript sources on which the reconstitution of Viana is based: the muster-rolls of 1826–1833, the electoral registers of 1834–1931, the passport books of 1835–1896, and the cemetery lists of 1855–1922.²⁷ Examples of these four sources are presented in Appendix A.

A range of other sources are also being used for the complete reconstitution of Viana: parish records of births, marriages, and deaths,²⁸ military recruitment registers, and notarial records, but their use lies beyond the scope of this study. It is worth noting here that although Portuguese confessional roles are known to exist, sometimes providing virtually continuous population registration, none have yet been discovered for Viana.

3.1 *The Muster-Rolls, 1826–1833*

Muster-rolls²⁹ have been drawn up in Portugal since the mid-seventeenth century, when Portugal regained her independence from Spain. Those drawn up

²⁷ Since the creation of each source was governed by various bodies of legislation, the indices compiled by Lencastre (1869) and Vasconcellos (1930) were consulted in order to identify important changes which could then be investigated in detail from primary sources. Except where otherwise indicated, references to legislation are from these sources.

²⁸ These are available after 1536 when the resolution that each church should keep a book of baptisms and deaths was passed (Alarcão, 1983: 22).

²⁹ In a previous paper (Kitts et al., 1987), the muster-rolls of Viana were referred to as militia lists. The current description is adopted to reflect more accurately their local nature.

between 1826 and 1833 were governed by the precise legislation of 1812,³⁰ and are the only systematically collected source of demographic data available for Portugal in the period 1820–1835 (Serrão, 1973 117). They are household listings – *Registos de Fogos e Moradores* (registers of hearths and residents) – each covering one Company of a military district, the limits of these districts corresponding closely with those of the *ancien régime* boroughs. Where more than one hearth existed in a house, each was distinguished by a letter of the alphabet.

The military District of Viana comprised 12 Companies: four in the Town itself, one including part of a road in Santa Maria Maior, the rural settlement of Abelheira in that parish, and the whole of the parish of Meadella (about 1km to the east of the Town, up the Lima valley), and the rest covering other surrounding parishes to the north of the River.³¹ The First Company covered the commercial centre of Santa Maria Maior, an area enclosed by the old town walls, the Third Company covered the area of Monserrate bordering the River, mainly comprising the fishing community, and the Second and Fourth Companies covered the rest of the Town, excluding part of the Rua da Bandeira which was included in the Fifth.

Unfortunately, under the legislation of 1812, those included on the rolls did not constitute the entire population, as spinsters and widows who were not heads of households and who did not have male children were not included. The rolls thus comprise: first, all males, second, females who could affect a male's liability for recruitment (wives, and single or widowed mothers), and third, since every household had to be registered, female heads of households where no males resided.

For each male inhabitant, name, marital status, date of birth (or age in years), birthplace, and father's name were consistently recorded, only name and marital status were generally recorded for females. Any further information relevant to potential military service was later entered in an observations column by the scribe. Such information typically included notes on an individual's availability for recruitment, changes of address both within and between military districts, departure elsewhere, and death.

These extra observations, together with information on individuals entering the list after its initial creation were written in different inks, by clerks with different styles of handwriting, often making it possible to distinguish between them.

30 Although the 1826–1833 muster-rolls were initiated just before, and updated during, the reign of Dom Miguel, there is no record to indicate that he was instrumental in their creation.

31 The other seven companies of the District of Viana covered the following parishes: Sixth Company, Areosa and Carreço, Seventh Company, Carreço and Afife, Eighth Company, Soutello, São Pedro and Âncora, Ninth Company, Vilar, São Lourenço and Amonde, Tenth Company, Nogueira and Cardiellos, Eleventh Company, Outeiro and Perre, and Twelfth Company, Santa Marta and Serreleis. Only Meixedo, Torre, and Vila Mou, the other three parishes of the *ancien régime* Borough of Viana are not included in any of the twelve companies.

3.2 The Electoral Registers

Electoral registers were first drawn up in Portugal according to the Decree of 9 January 1834, this was passed by the liberals more than four months before the final surrender of Dom Miguel. The legislation stipulated that elections for members of a *Câmara* (municipal chamber)³² were to be held annually in every borough. The electoral registers themselves were to be drawn up by parish, differentiating between those only eligible to vote, and those also eligible to be elected. They were revised annually until 1878, but less frequently thereafter. The registers of Santa Maria Maior and Monserrate for the following 54 years are considered here: 1834, 1835, , 1878, 1880, 1881, 1883, 1888, 1891, 1894, 1895, 1911, and 1931.

A major problem immediately arises in the use of these registers, however, since an individual's appearance on them depended on political rather than demographic considerations. In order to identify the electorate, and more particularly, its relationship with the population from which it was drawn,³³ it is necessary to review the legislation governing the creation of these data.

Fortunately, membership of the electorate depended little on social groupings such as class or occupation, that as a rule mattered most, but rather to the observance of rules applied nationwide whose relevance to the local community could be minimal (Feijó, 1983 293–294). However, electoral legislation was nevertheless changed frequently between 1834 and 1851 amidst the political turbulence of the first phase of Portuguese liberalism, and it was only after this latter date that a certain stability in both the political climate and the electorate on which it relied is deemed to have emerged (Feijó, 1983 301).

Essentially, the electoral legislation of nineteenth century Portugal, like that of many other countries, was based upon the idea that the 'active citizen' was the foundation of the state. Those who triumphed in 1820, when the first version of the constitution was proclaimed, and who remained victorious after 1834, were careful to define the elector as an independent, settled man of property (Silbert, 1977)³⁴.

While the earliest electoral legislation simply stated that an elector should have an income of 200\$000 *reis* derived from landed property, commercial activity or employment (Constitution of 1826), subsequent measures provided for the detailed computation of a man's income from a variety of fiscal documents. The Law of 27 October 1840, stipulated that the commission charged with the compilation of the electoral register should consult the tax records, to verify that an elector had made a payment of 10\$000 *reis decima* on incomes received from the local authorities or the *Casa da Misericórdia*, or had paid a tax of 5\$000 on property rents owed to him, or 1\$000 *reis* property tax on his own residence,

32 Although every borough had a *Câmara*, the process of formation and the attributes of each varied enormously, what a *Câmara* could and could not do was a combination of legal disposition and custom (Feijó, 1983 294).

33 Brandão & Feijó (1984) warn the researcher of the 'truncated image of a society subject to artificial divisions which altered a lot over time' offered by these registers.

34 The Law of 9 January 1834, provided that elections be conducted according to the instructions set out in the Constitution of 1826.

or on income derived from his own enterprise

In addition, stipulations that the level of cash should be earned in a consistent manner by a resident of long standing were designed to keep the electorate free of the urban journeymen whose activities in times of crisis in 1836 and 1838 had proved a potent threat to the regime.³⁵ An elector's residence, defined as 'that part of the country where he resided for the greater part of the year', also provided against the possibility that politicians might influence the outcome of elections by the judicious mobilisation of state officials.³⁶

The administrative reforms of the 1840s also provided for the 40 largest taxpayers of the Borough to play a consultative role in local affairs – the committee of scrutineers who met every year (usually in July) to draw up the electoral register.³⁷ These scrutineers possessed the right to inspect the qualifications of those who did not produce evidence from the Town's fiscal records, and were responsible for verifying whether other 'proofs of income' were acceptable equivalents to the tax returns.³⁸ They were also empowered to exclude those who were too young and did not possess the appropriate educational qualifications, those who worked in personal or domestic service, and freedmen, regardless of wealth.

The 40 largest tax payers continued to play a major role in the formation of the electoral rolls throughout the 1850s and 1860s, buttressed by their preeminence in the *Câmara*.³⁹ Their position was retained in the bodies of legislation introduced in the 1850s. The financial calculations were amended in 1852 to allow a prospective elector to demonstrate that he possessed sufficient income by having paid 1\$000 *reis* in the new 4% property tax levied on personal residences.⁴⁰ These tax considerations were further modified with the progressive legislation of 1859, which provided for the inclusion of two new taxes, the *decima industrial* and *contribuição predial*.⁴¹

Only after 1878 was the electorate broadened to include literate heads of households, and income qualifications abolished,⁴² a process completed by 1883.⁴³ Also, by this time the registers were compiled by the municipal authorities, without interference from scrutineers.

In 1895 and 1896, however, the franchise was again restricted, so that only

35 Urban journeymen had played a major role in the rising of September 1836 (Peres, 1935)

36 Law of 27 October 1840 (*Legislação Portuguesa*, 1840)

37 Administrative Code of 18 March 1842 (*Legislação Portuguesa*, 1842)

38 Members of the regular clergy, graduates of Coimbra University, those with formal qualifications from another university, polytechnic, or lycée, and teachers in secondary schools or higher education were dispensed such proofs of income, subject to the scrutiny of the committee. Law of 30 September 1852, Art 7, III-VIII (*Legislação Portuguesa*, 1852)

39 The restrictions on eligibility for election, far stricter than those governing the right to vote, ensured that the *Câmara* was essentially composed of members of the local aristocracy and businessmen, and not 'the people' (Feijó, 1983 298)

40 Law of 30 September 1852 (*Legislação Portuguesa*, 1852)

41 Law of 23 November 1859, Art 2 (*Legislação Portuguesa*, 1859)

42 Law of 8 May 1878 (*Legislação Portuguesa*, 1878)

43 Law of 4 April 1883 (*Legislação Portuguesa*, 1883)

those who could read and write, and were paying at least 500 *reis* directly to the state qualified to be electors automatically, while anyone meeting the literacy requirements but not the financial ones merely had the right to petition personally for the right to vote ⁴⁴

Before the proclamation of the Portuguese Republic in 1910, there was a further widening of the electorate in 1899, and a narrowing in 1901 ⁴⁵

After the proclamation of the Republic, the electorate was broadened to include all Portuguese citizens over the age of 21 who could read and write, or who were heads of a households ⁴⁶

By 1931, the most recent year for which electoral registers are considered, this legislation had been modified such that those who could not read and write were to have paid at least 100\$00 escudos in taxes, also, females over the age of 21 with certain qualifications were allowed to vote ⁴⁷ However, the register of 1931 must be approached with some caution, because it was compiled under dictatorial rule ⁴⁸

In summary, the most important bodies of legislation are shown in Table 2, classified according to whether they were progressive – widening the franchise – or regressive – narrowing the franchise

Table 2. Portuguese Nineteenth-Century Electoral Legislation

Date	Effect	Government
7 8 1826	Regressive	Trigoso Morato
4 6 1836	Regressive	Duque de Terceira
8 10 1836	Progressive	Passos Manuel
5 3 1842	Regressive	Costa Cabral
30 09 1852	Progressive	Duque de Saldanha
23 11 1859	Progressive	Duque de Terceira
8 5 1878	Progressive	Rodrigues Sampaio
28 3 1895	Regressive	Hintze Ribeiro, João Franco
21 5 1896	Regressive	Hintze Ribeiro, João Franco
26 7 1899	Progressive	José Luciano de Castro
8 8 1901	Regressive	Hintze Ribeiro

Source Serrão (1976)

It is noted that two important bodies of legislation do not appear in the table

44 Laws of 28 March 1895, and 21 May 1896 (*Legislação Portuguesa*, 1895, 1896)

45 Laws of 26 July 1899, and 8 August 1901 (*Legislação Portuguesa*, 1899, 1901)

46 Law of 14 March 1911 (*Legislação Portuguesa*, 1911)

47 Laws of 2 March 1928, and 27 December 1933 (Godinho, 1969)

48 General Carmona carried out the military *coup* of 1926 By 1929, he was giving way to his Minister of Finance – António de Oliveira Salazar (Thomson, 1962 633)

First, the progressive constitution proclaimed after the Revolution of 1820, under which Portugal held its first parliamentary elections. Second, the progressive decree of 1834, which provided for regular parliamentary elections and the compilation of the necessary electoral registers

Finally, with regard to the information recorded on the electoral registers, it is important to bear in mind that this varied from year to year according to the legislation. Name, age, marital status and occupation are recorded on almost every register,⁴⁹ together with one or more indicators of income (either direct or indirect). Road of residence is recorded on the first three registers, while birthplace is recorded on four subsequent registers. A number of other observations such as literacy, whether the elector was the head of a household, whether he was eligible to be elected, whether he had performed jury service, etc., appear at various times during the period. A detailed description of the data recorded on the electoral registers is presented in Appendix B.

3.3 The Passport Books

Legislation passed in 1835 provided that passports for travel outside Portugal were to be issued by the Civil Administration of each district at a price of 800 *reis*, the particulars of each passport issued being recorded in passport books. The District Administration of Viana do Minho issued its first passport under this legislation on 8 October 1835.

The first passports carried the name of the holder, his birthplace, marital status, and occupation, his destination, the name of the individual declaring himself responsible for the application (this item was dropped during the 1850s), and a number of descriptive particulars concerning age, height, face, hair, eyebrows, eyes, nose, mouth, colour, and any other distinguishing features. In addition, the date of issue was recorded, together with the length of time for which the passport was valid (usually 60 days), and the signature of the official issuing the passport. Space was also provided for the recording of the intended travel itinerary (port of embarkation, etc.), but this was rarely used. This basic structure soon evolved to include the passport holder's current place of residence, his father's name, space for the inclusion of information regarding individuals accompanying him. Much later, in 1927, passport photographs began to appear.

Only the most important items of information are considered here: birthplace, sex, name, age, marital status, occupation, date of issue, and destination. These are available for the 1,854 people who appear on the 1,522 passports issued to residents of the Town/City of Viana before 1896, when changes in the recording of information made it difficult to accurately ascertain current place of residence.

3.4 The Cemetery Lists

The repeatedly delayed adoption of public cemeteries in nineteenth century Portugal lies rooted in the population's reluctance to abandon traditional burial

⁴⁹ Only names were recorded on the registers of 1837–1839, and marital status was also omitted from the registers of 1840.

practices, such as within the walls of churches. Attitudes to death and burial in the Minho are discussed in detail by Pina-Cabral (1986 214–226), and their effect on the introduction of public cemeteries in the Borough of Viana is discussed by Feijó (1983 316–322).

Although the first national bill on the subject was passed in 1835,⁵⁰ it was not until December 1840, that the first public cemetery was opened in Viana, and then the public discontent and careless manner in which it was maintained led the Civil Governor to order its closure just eighteen months later.

This slow progress led the Government of Costa Cabral to pass the Health Laws of 1844 which sought to tighten control over burial practices by including measures, amongst others, forbidding burial in churchyards and ordering them to be interred at some distance from villages. However, as mentioned in Section 2.2, violent protest was precipitated, and some authors (e.g. Nowell, 1952 196) go so far as to suggest that the revolution of Maria da Fonte was caused principally by peasant reaction to these laws.

In fact, it was only in 1855, when cholera threatened the City of Viana killing 95 people in the parish of Santa Maria Maior alone, that a new public cemetery was inaugurated (Castro, 1954, 1955). The first entry in the cemetery lists of Viana was made on 24 September 1855.

The lists record the name, father's name, birthplace, marital status, occupation, date and place of death of the deceased, together with the date on which they were buried and a number identifying where they were in the cemetery.

Considering the use of the lists, two caveats arise. First, information concerning age at death was initially omitted, thus exacerbating the difficulties of reconstitution. This problem is alleviated, albeit only slightly, from 1863, when separate books began to be kept for minors (aged less than 14 years) and adults (aged at least 14 years), and removed completely from 1879, when age at death was included. Second, the census of 1862 specifies four places in which people could be put to rest, namely plots in public cemeteries, private vaults in public cemeteries, private vaults elsewhere, and the church. However, those who chose the latter two were by this time mainly monks, nuns, and important aristocrats.

In view of these drawbacks, it is necessary to justify the use of the cemetery lists rather than the death records which probably cover all deaths, include information on age at death, and are available for each parish separately throughout the period on which this study focuses. First, the cemetery lists have the practical advantage of actually being lists (i.e. information on each individual is arranged in columns) and being available in photocopied form, while the death records are less structured and are only held on microfilm. Second, it is believed that the items of information available on the lists combined with the knowledge

50 The Decree of 20 September 1835, stipulated that cemeteries were to be provided (Art 1), they were to be built outside city, town, or village limits (Art 3), that they were to be surrounded by a wall about 10 hands high (Art 4), and that a separate grave was to be provided for each corpse (Art 5). The Decree of 8 October 1835, further ordered that parish and borough authorities were to be responsible for the upkeep of cemeteries (Art 1), and that the corpses of people who had had an annual income of less than 100,000 *reis*, or who had not been electors, were to be buried at no cost (Art 3).

that an entry thereon ought not to precede any entries on the electoral registers or in the passport books are sufficient to minimise the likelihood of incorrect linkage with records of the early lists on which age did not appear. Finally, the number of deaths for which death records but no entry on the cemetery lists exist is believed to be negligible since a comparison of several years of both sources yielded little disparity.⁵¹

4. RECORD LINKAGE

In philosophy, great distinction is drawn between knowledge by acquaintance and knowledge by description, where the importance of the latter is that it enables people to pass beyond the limits of their private experience. Russell (1912: 31) gives the example that given the knowledge that 'Bismarck was an astute diplomatist', one would like to affirm, when describing Bismarck as the first Chancellor of the German Empire, that 'the first Chancellor of the German Empire was an astute diplomatist'. This process involves the acceptance that the objects referred to (Bismarck) are in fact the same, and therefore constitutes an increase in knowledge using record linkage. In this way, most knowledge is by description, and in particular, historical knowledge is entirely by description, leading Winchester (1972) to write that 'History is speculation about the past controlled by record linkage'.

Record linkage then, although often not identified as such, is a fundamental human process that has long been applied. The twentieth century, however, has seen its development as a technique for the identification of large numbers of persons from any number of different types of records, and its application in various fields including commerce, demography, genetics, and medicine. In historical demography, record linkage is most commonly associated with the family-reconstitution techniques which, although often attributed to the pioneering efforts of Marcel Fleury and Louis Henry in the 1950s, were applied as early as the 1910s in Sweden (Åkerman, 1982: 187). More recently, the availability of larger and more powerful computers make fully automatic record linkage possible. This has encouraged the development of many large scale projects, however, the majority of these remain overshadowed by the work of the Demographic Data Base at Umeå University, Sweden.⁵² Thus, in the last three decades, the need for formalisation of techniques has led to the generation of a wide, varied, and interesting body of literature on the subject of record linkage.

⁵¹ By the same token, of course, this suggests that the number of deaths not recorded on the parish registers of Viana is small.

⁵² See Danell (1981), or Sundin (1984) for a description of the Demographic Data Base at Umeå.

Reviews of this literature are provided by Winchester (1973b, 1974, 1985)

This Section reviews current practices and introduces recently-developed software which facilitates the record linkage process. First, a brief review of the literature focuses on the problems of record linkage in historical demography in order to identify the software requirements of record linkage. Second, the Scientific Information Retrieval (SIR) Data Base Management System (DBMS) is described, and its ability to meet the software requirements of record linkage is demonstrated. Finally, record linkage using SIR is described in more detail.

4.1 Issues in Record Linkage

Record linkage might initially be defined as the process whereby pairs of records, one from each of two files of records, are compared in order to ascertain whether they both relate to the same person, if there is sufficient agreement between the records along some predetermined criteria, the records are linked, otherwise they are not. This definition, however, becomes unsatisfactory when two or more records in one file contain identical sets of information, or more than two files of records are being considered. In the former case, it is usually impossible for two records in one file to refer to the same person. In the latter case, given three records, A, B, and C for example, A may be linked to B, whereupon consideration of C reveals that while B and C ought to be linked, A should not be linked to this pair. Thus, record linkage comprises the problem of deciding whether two records might match, and where more than one possible match exists, the problem of deciding which match is better. This naturally extends to higher order problems where clusters of possible matches exist. Increases in dimensionality are discussed later, but first attention is restricted to comparisons of pairs of files of records.

Having defined record linkage as above, the problem reduces to one of calculating some measure of consistency between any two records under consideration. Further, since each record comprises an unspecified number of identifying items, or fields, such a measure can be calculated from similar measures of consistency between each pair of fields. This approach is illustrated with two trivial examples. On the one hand, if the sex fields of two records are compared, then either they are the same, in which case they are completely consistent, or they are not, in which case they are completely inconsistent and the comparison of the two records can terminate. On the other hand however, consider a comparison of the age fields of two records. If age was recorded as the number of complete years lived, and the interval between the compilation of the records was not an integral number of years, then the two recorded ages, while both being accurate, are not unlikely to differ by one year.⁵³ Further of course, there are many other reasons why two fields may not be fully consistent: misreporting, misrecording, variation in spelling, non-contradictory description, change, and even errors introduced by the researcher when transcribing the data. This

⁵³ For example, an individual recorded as being 10 years old on 31 December 1986 must have been born in 1976. However, if his age is recorded again on 31 June 1987, it may be 10 or 11 years.

introduces the most important issue in record linkage in historical demography, that 'by meta-Leibnizian principles, a pair of identifying items which on the surface seem to be contradictory are, at a deeper level, equivalent' (Winchester, 1972 7)

Some examples of surface differences are illustrated with the Berkeley/Barclay puzzle (Winchester, 1973a 32)

Surname	Initials	Birthdate	Birthplace	Occupation
Berkeley	G J	1676	Dublin	Sailor
Barclay	J G	1667	Ireland	Tailor

If these two records are linked, the puzzle becomes one of whom is identified? Berkeley or Barclay? And what is one entitled to say about him?

With textual data in particular, standardisation and coding of the data can be used to remove many surface differences, thus simplifying the comparison of fields, this essentially amounts to treating certain pairs of words and phrases as fully consistent. In historical demography, the majority of these surface differences arise because certain pairs of letters and digraphs either sound the same, or look the same. First, the non-standardisation of orthography in earlier periods caused variations in the spelling of many words, this is particularly problematic in the comparison of names. Since the written form was derived from the spoken form, these variations are often easy to identify and have been approached using phonetic coding systems which attempt to reverse the name recording process, so that different written forms of the same spoken word are coded identically. To this end, many researchers have adopted the original Russell Soundex code, some modification thereof, or other similar systems (Newcombe et al, 1959, Newcombe & Kennedy, 1962, Newcombe, 1967; Nitzberg, 1968, Smith, 1968, Winchester, 1970, Légaré, 1972, Blayo, 1973; Pouyez et al, 1983)⁵⁴. However, success with historical material has remained somewhat limited (Wrigley & Schofield, 1973 98). Second, there is the problem that certain pairs of letters are very similar in their script form, as for example the S/T in the occupation field of the Berkeley/Barclay puzzle. This problem has been approached using a Viewex code (Winchester, 1968). Ideally of course, a coding system incorporating a combination of Soundex and Viewex is desirable (Winchester, 1970).

⁵⁴ The Russell Soundex code, for example, is derived as follows. 1 The initial letter of the family name is used as such without numerical code and serves as a prefix letter. 2 Letters w and h are always disregarded except when serving as prefix letters. 3 Vowels a,e,i,o,u, and y are not coded, they serve as dividers (c f rule 5 below). 4 The following letters, not to exceed three in number, are coded as follows (subsequent letters are dropped) b,p,f,v=1, d,t=3, l=4, m,n=5, r=6, all other consonants (c,g,j,k,q,s,x,z) are coded 2. 5 Exceptions to the above rule are those letters after the prefix calling for the same code number, unless they are separated by a divider (c f rule 3 above), the second is dropped. Source Légaré (1972 438).

Considering names, even once orthographic variations have been removed, several problems, often peculiar to the particular society under study, remain to be tackled. In Sweden, during the nineteenth century, it was common for people when adult to abandon their patronymic name and take a new surname (Sundin, 1985). In late medieval and early renaissance Italy, names were often shortened or lengthened (e.g. Vestro for Silvestro, or Marchetto for Marco) (Herlihy, 1973: 49), and inversion of name order was not uncommon (Skolnick, 1970). In the linkage of census name data from the Saguanay region in Quebec, between 1850 and 1861, these sorts of problems cause Pouyez et al (1983) to find that men's names are consistent from one census list to the next only in 78% of the cases. This might be compared with record linkage at the Demographic Data Base at Umeå University, Sweden, where more than 90% of entries can be linked using only parish of birth and date of birth information (Sundin, 1984). Perhaps the most sophisticated programs for handling names that have been developed so far are Winchester's. These involve prefix, infix, and postfix treatment of family names, combined with a Viewex/Soundex coding scheme. These transformations and the associated program are reviewed by the Pennsylvania Social History Project (Hershberg et al, 1976).

Considering fields other than name. Textual data suffer similar spelling variation, and, in addition, sometimes two different descriptions of a field need to be treated as fully consistent (e.g. place or occupation). For numerical data, Pouyez et al (1983) find that age discrepancies in truly linked record pairs ranged from zero to fifteen or more and were most frequently around 8–10 years, although discrepancies of fifteen or more years were by no means insignificant (6% of cases). Discussion of other fields and their treatment with respect to the problems of non-contradictory description and change can be found in Newcombe et al (1959), Newcombe & Kennedy (1962), Wrigley (1966), Nitzberg (1968), Tepping (1968), Winchester (1970), Katz & Tiller (1972), Hershberg et al (1976) and Pouyez et al (1983).

Having reviewed the standardisation and coding of data, attention is returned to the comparison of two records, one from each of two files. Considering the total number of possible links, where two files contain r_1 and r_2 records, it has hitherto been assumed that $r_1 r_2$ potential links are to be examined. However, this total can be dramatically reduced if each file is subdivided into blocks, according to the value of some highly reliable variable. For example, if each file is subdivided by sex, such that $r_1 = rm_1 + rf_1$ and $r_2 = rm_2 + rf_2$ (where rm and rf are the numbers of male and female records, respectively), then the total number of comparisons can be substantially reduced to $(rm_1 \bullet rm_2) + (rf_1 \bullet rf_2)$. Of course, the reduction in the total number of possible links increases rapidly with the number of values that the subdivision variable can take. Such methods of sorting and merging files, which optimise the search for reasonably comparable records, are discussed by Iverson (1962), Nathan (1964, 1967), and Newcombe (1967). It must be emphasised however, that the subdivision of two files into blocks is actually part of the record linkage process, in that it precludes linkage between blocks.

Indeed, in situations where the data are of fairly high quality and record linkage is relatively straightforward, similar non-probabilist methods, perhaps

augmented by a simple set of tolerance rules, have been used for the entire record linkage process (Phillips et al, 1962, Hubbard & Fisher, 1968, Légaré, et al, 1972, Pouyez et al, 1983, Bouchard 1986)

In more complex situations, the comparison of two records leads to what Winchester (1974:36) describes as the 'woolly regions of weighting systems'. Leaving aside *ad hoc* procedures, three main approaches are reported in the literature. First, and perhaps most important, Felligi & Sunter (1967, 1969) and Sunter (1968) offer a formal mathematical description of Newcombe & Kennedy's (1962) work on record linkage, they propose a linkage strategy which seeks to minimise the number of records for which no linkage decision is made – under the assumption that the probabilities of erroneous matches and non-matches are fixed in advance. Second, Du Bois (1965, 1969) attempts to maximise the number of true links while simultaneously minimising the number of false links. Third, Nathan (1967) and Tepping (1968) attempt to minimise the expected cost of assigning links on the basis of a predefined comparison function (Winchester, 1974 52–55).

From a statistical point of view the most intuitive of these three methods is the first – a likelihood ratio approach. This is essentially as follows. Assuming that the fields of records are mutually independent,⁵⁵ the likelihood ratio, or odds, of two records referring to the same historical person is calculated as the product of the likelihood ratio of each pair of fields. These ratios are a measure, then, of the probability that a particular field is consistent (or inconsistent) given that they are linked. In practice, the logarithm to the base 2 of each ratio is usually calculated, thus producing additive weights – binit weights (Newcombe & Kennedy, 1962).⁵⁶ However, this approach requires that those ratios be known in advance, which in general they are not. They are usually calculated from a file of truly linked records as the number of truly linked pairs for which the combination of two fields arises, divided by the number of falsely linked pairs for which the combination arises. Unfortunately, this method of estimation of the probabilities can itself introduce bias and intercorrelations between fields (Wrigley & Schofield, 1973:93). Once all possible links have been examined, and their likelihoods calculated, the file of links can be sorted in decreasing order of likelihood, and the data linked from the most likely to the least likely pair above some predetermined likelihood. Even having adopted this approach, however, Hershberg et al (1976:161) remark that their file of linked persons is biased 'toward individuals with characteristics which are uncommon, particularly toward those individuals with uncommon names'.

To complete the review of record linkage, situations where there is duplication of identifying item sets, or there are more than two files of records must be discussed. The important difference between these and the previous situation is that the total number of possible solutions increases rapidly with

55 Intercorrelations between certain fields, such as age and marital status for example, do of course exist. However, where comparisons are based on the same fields, the resultant likelihoods will nevertheless comprise an ordinal scale, and can still be used meaningfully to prefer higher-scoring links (Wrigley & Schofield, 1973 93–94)

56 A useful account of this procedure can be found in Wrigley & Schofield (1973 92)

the number of duplications, or additional files of records,⁵⁷ leading to 'the even woollier thickets of investigating networks of quasi-linked records' (Winchester, 1974 36) The processes involved here lean towards statistical cluster analysis, where some measure of the similarity between any pair of records is defined as before, and, according to this measure the records are gathered into groups or networks within which at least one of the links between each record and any other record of the same group is associated with a measure that exceeds some predetermined level This procedure is then followed by one of examining each cluster in order to subdivide it into a number of historical persons while simultaneously maximising some measure of the likelihood that these subdivisions are indeed the correct ones Methods for resolving such clusters are discussed by Skolnick (1973), Wrigley and Schofield (1973), and Bouchard (1986)

In conclusion, record linkage studies can be classified according to whether or not variation or errors or both exist in identifying items of information, and whether or not there is duplication of identifying item sets (Winchester, 1973b, 1974, 1985) In the case where neither problem exists, record linkage becomes a trivial exercise, for example, this arises in the linkage of elites (e.g. Drake, 1971) Otherwise, some measure of the consistency of pairs of fields and then pairs of records is required, and a strategy for comparing clusters of possible links may also be required

At the same time, although several general approaches and methodologies for efficient record linkage exist, each new study will present its own peculiarities, and will require its own particular set of priorities in identifying two separate records as pertaining to the same historical person The type of data recorded, the frequency of recording, and the way in which data were recorded will influence the researcher's choice of algorithm to resolve ambiguities in record linkage, while maximising the accuracy of the constructed links Nevertheless, whatever approach is adopted ought to be fully automatic, ensuring both that linkage criteria are carefully defined beforehand, and that those criteria are consistently applied

Also, it can be seen from the above discussion that in order to enable records to be linked both within and between documents for subsequent analysis, several operations – the software requirements of record linkage – must be possible First, and foremost, the data must be easily accessible for editing and, later, analysis Second, it must be possible to transform or code the data in order to eliminate the simpler types of variation, such as spelling Third, the comparison of one or more variables from each record should be a simple process, in order to allow concentration on the more complicated aspects of record linkage Finally, it must be possible to store the links produced so that the linked data can be efficiently retrieved for subsequent analysis One way to satisfy these requirements is to store the data in a powerful database In Section 4.2, reasons for using the Scientific Information Retrieval (SIR) Data Base Management System (DBMS) for the storage, linkage, and retrieval of the Viana data are discussed

57 The combinatorial problem of record linkage is formally illustrated by Kelley et al (1972), and Skolnick (1973)

4.2 The Scientific Information Retrieval Data Base Management System

SIR is a hierarchical database system in its physical form, each record is stored following the record to which it relates. Thus, the set of information relating to each individual-entry on a document is given a unique identifier – a CASE in SIR – which points to several RECORDS concerning that entry, each of which, in turn, may contain several FIELDS (or VARIABLES) ⁵⁸ Because of the uncertainty inherent in historical demographic data, it would make little sense to use one of the fields of information of an individual-entry as the case identifier. Instead, a number for the individual-entry (NID) can be generated which may incorporate digits representing the document from which the individual-entry is drawn, the date it was created, its position within the document, etc. A major advantage of SIR is that, although it has a hierarchical physical model, one is not restricted to hierarchical access, the logical model of the database allows hierarchical, relational and network access. It is the network facility that is most useful for record linkage, since from any position in the hierarchical physical model it is possible to re-enter the hierarchy at any other whilst retaining the original position for subsequent continued search.

The SIR system has been shown to be extremely useful in the development of a general approach to the machine handling of event history data (Ní Bhrolcháin & Timaues, 1983). It is a natural extension then, for it to be used in the handling of record linkage material which comprises events that are to be linked into person, or individual life histories.

The biggest problem in the computer-handling of manuscript sources is that the bulk of them contain textual variables. It soon becomes apparent that, with each name, occupation, birthplace, etc., appearing many times on various documents, the space saved by coding each variable would be enormous. Lookup tables can be created as records in SIR by constructing dummy cases containing the coding tables for any number of textual variables, so that only numbers are recorded in the fields of each case. Also, similar coding tables can be used to store standardised versions of textual variables, and associated indicator variables which might be required for record linkage or subsequent analysis. Further, this coding and standardisation of data, which is essentially a simple extension of the way in which marital status is often recorded (i.e. Single = 1, Married = 2, Widowed = 3, etc.), simultaneously enables the frequencies of occurrence of each variable to be stored for use in record linkage programs, and allows the data to be processed more efficiently because numerical values for fields are available.

In order to illustrate these facilities, the storage of names in the Viana Database is described. When a name is read into the database from a raw data file, it is split up into its component names – a title (if any), a first name, and other, second names. For each component name, existing codes are used, or

⁵⁸ Some slight confusion is unavoidable here because the term record has a different meaning in SIR than it has in the record linkage literature. The CASE in SIR often represents the record discussed earlier; the disparity arises because of the intermediate level of RECORD in SIR, into which closely related FIELDS (or VARIABLES) are grouped.

new codes are created, these codes are stored in the relevant fields of a name record belonging to that case. At the same time, frequencies of occurrence are recorded, and the codes are indexed so that each component name points to all the cases which contain it in their name record. Next, a reverse lookup table record is created so that the original component name can be retrieved from the recorded code. This lookup table also contains a standardisation variable into which the code of a standardised version of the component name is entered. The standardisation variable can be generated automatically, or entered manually. If it is entered manually, standardisation is initially quite laborious, but offers the enormous advantage of handling any type of data coding and standardisation. An essentially similar, but slightly simpler, procedure is used to store textual information on occupation, and birthplace, where the raw data are not split up into individual words.

Finally, an additional advantage of the system described here is that while data are stored in their original form, and can be retrieved as such, standardised data can be used for micro-analyses, thereby offering researchers the option of being spared confusion arising from variations in spelling, etc.

4.3 Record Linkage using the SIR DBMS

In SIR, the number uniquely identifying each case, or individual-entry – the NID – can be used to generate links which are stored in a subordinate record for each case. This record contains the NID of the first individual-entry (FID), the NID of the previous individual-entry (PID), and the NID of the next individual-entry (LID) of each identified historical person. The LID of the last individual-entry is set to equal the FID, so that a chain is formed for each person, the end of which points back to its beginning. Subsequent identification of relationships between different historical persons can then be achieved by linking two or more chains together. For example, a family link record may be created to contain those links required for family reconstitution, one of the variables in this record might contain the FID of the father of a person.

The FID serves several purposes. First, it enables a chain to be followed from its beginning even if it was entered at some other point. Second, it is possible to determine when the end of a chain has been reached. Third, when processing all the cases in the database sequentially, and following the chain of individual-entries for each historical person at the same time, it enables SIR to determine whether the information on one person has already been processed. Finally, the FID is the identifier used for linking different chains, as described in the previous paragraph.

Thus, the versatility of SIR allows attention to be focused on the major problem of record linkage exercises – the determination of a set of linkage criteria, or control variable which will maximise the accuracy of the links generated. The choice of control variable will depend primarily on whether the researcher chooses an agglomerative or a divisive approach to record linkage, or a combination of the two.

An agglomerative approach involves a reduction of the number of variables

contributing to the control variable, thus relaxing the constraints under which a link is made, in this way, groups of records belonging to persons can be brought together. This method is particularly applicable when there is a limit to the number of records which can belong to one person, it is therefore being used in the linkage of birth, marriage, and death records. It is possible to control the generation of links, so that previous links which were generated using stronger constraints, are not overwritten.

A divisive approach involves an increase of the number of variables contributing to the control variable, thus tightening the constraints under which a link is made. Usually, this approach will begin with just one variable, the standardised full name, so that all the records which might belong to one or more persons can be brought together. The resultant links can then be examined, and any records which are inconsistent under a specified set of criteria, can be removed and placed into other chains. This method is particularly applicable when there may be many records belonging to a person, but few persons share the same initial variable.

The operations of joining and separating different chains are fairly complex procedures. In order to join two chains which are non-overlapping, the end of the first can simply be made to point to the beginning of the second. Otherwise however, one of the chains has to be completely dismantled, and its individual members inserted into the other chain. Similarly, to divide a chain into two non-overlapping sub-chains is quite straightforward, whereas to extract selected members from one chain to form a new chain is more complicated.

5. PREPARATION OF THE VIANA DATA

This Section illustrates and discusses the technical problems arising in the use of both the textual and numerical data recorded on the manuscript sources of Viana, and the methods and techniques which have been adopted to overcome them.

5.1 Data Entry and Storage

In order to retain the data in their original form without altering the structure of the original manuscript sources, and to allow any item of information to be traced back to its source document, each set of information relating to one individual on a particular document (individual-entry) is given a unique identifier, the number of the individual-entry (NID) comprising eight digits: the first coded according to the nature of the document,⁵⁹ the next three to the year in which it was drawn

⁵⁹ Document code 1 – Muster-Roll, 2 – Electoral Register, 3 – Passport Book, 4 – Cemetery List

up, the fifth to the parish to which it related⁶⁰ and the last three to the position of the information within the document⁶¹ This coding is such that if data are sorted by NID, they are automatically sorted first within a document, then by parish, next in time, and finally by type of document For example, to retrieve the standardised full name of the first individual recorded on the 1840 electoral register of Santa Maria Maior, the case 2-840-1-001 is located, and then the name record it points to From this record the name code variable is extracted, and used to retrieve the individually standardised component names from the name lookup table record, these are then concatenated to form the standardised full name

Initial experiments used a BASIC input program which could be modified for each type of document Run on the BBC microcomputer, the program asked for the NID of the first individual-entry, and then asked for every possible item of information relating to consecutive individual-entries However, it was found to be far more efficient for the data on documents in columnar format to be entered one variable at a time Two methods were therefore adopted for the input of textual data More experienced researchers entered data directly into ASCII data files using a variety of word processing packages on the BBC, IBM-PC-AT and IBM-PS/2 microcomputers Those with less experience entered data into mini SIR databases on IBM PC-ATs which built up validation tables so that when a component name not already known to SIR was entered the researcher could seek confirmation before continuing On the other hand, numerical data were always entered directly into ASCII data files Only information on marital status was coded when the data were originally entered

Currently, there are approximately 7,000 muster-roll, 39,000 electoral register, 2,000 passport book, and 12,000 cemetery list individual-entries stored on the University of Southampton mainframe computer – an IBM 3090-150

5.2 Problems of the Manuscript Sources

Considering textual data, first, the orthography of Portuguese had not yet been standardised in the nineteenth century, and certain characters were sometimes interchangeable,⁶² and others, usually silent, were often omitted entirely⁶³ Second, the documents were often written in extremely abbreviated form, such that words would sometimes be truncated to the first letter followed by a full stop Third, where the document had a fixed columnar format, scribes would often

60 Parish code 0 – Viana (Town/City), 1 – Santa Maria Maior, 2 – Monserrate Only the electoral registers were drawn up separately for each parish, the code for Viana (0) is used for the other documents

61 For the muster-rolls, the NID is modified to comprise the document type, and address of the individual, including the road, house, and position of the information within that house

62 For example, the B and V (and W in the nineteenth century) are sometimes interchangeable, this peculiarity of pronunciation is still common in the Minho Similarly, the letter x is pronounced the same as the digraph ch, with which it was therefore sometimes interchangeable

63 For example, the C, M, and P, are sometimes silent, as in Victorino/Vitorino, Sam Miguel/São Miguel, and *esculptor/escultor*, respectively

reduce their work by using the word *idem*, or just quotation marks,⁶⁴ to indicate that an item of information was the same as that for the previous entry. Finally, the obvious difficulties of reading early nineteenth century Portuguese script must not be under-estimated, it is extremely difficult to differentiate between certain combinations of hastily scrawled letters, even with a thorough knowledge of the language.⁶⁵

Names, the most powerful of the variables available for record linkage, not only present the difficulties of textual data outlined above, but also several of their own. First, Portuguese names have been abbreviated for centuries according to unwritten convention (some of the oldest abbreviations can still be found in common use). Unfortunately however, although some researchers have suggested that every abbreviation can be found in the lists of any modern Portuguese telephone directory (e.g. Rowland, 1987), relatively few of those encountered in nineteenth century manuscripts actually appear, and where they do, the abbreviations sometimes generate further ambiguity.⁶⁶ Under these conditions, record linkage must be a recursive process, alternating between exploratory linkage and the subsequent 'decoding' of hitherto unknown abbreviations. Second, the formation of Portuguese names does not follow the relatively predictable pattern of northern European countries. They are essentially composed of one or more first names followed by one or more second names,⁶⁷ any of which might be passed on to the next generation.⁶⁸ The flexibility thereby introduced not only exacerbates linkage between generations, but also the linkage between different entries of a person since it was not uncommon for someone to change his name, either by dropping a component name, merging two component names,⁶⁹ adding a component name, or even just altering the order of two or more component names. Also, since first and second names are often passed on in twos or threes, it cannot be assumed that the probabilities of component names occurring in a full name are statistically independent.⁷⁰ A further result of the name formation process is that it was not uncommon for an individual to have a total of four or

64 In fact, it appears that quotation marks were used on the early electoral registers to indicate that a particular item of information was not available, and were only used in the same sense as *idem* after 1840.

65 Examples include B/R, F/T, L/P/S, T/V, A/M, a/o, g/q/z, l/t, n/u/v, in/n/m, m/π, etc.

66 For example, Martins was abbreviated to Miz, it is now abbreviated to Mart.

67 The term second name is used to avoid confusion with surname – a hereditary family name transmitted in male line.

68 See Feijó (1987) for an exploratory analysis of name formation using the muster-rolls of the parish of Carreço (just north of Viana).

69 For example, while it is simple to separate Villas-Boas into the two names Villas and Boas, the occasional merging of a d' with a name beginning with a vowel (e.g. d'Antas/Dantas) is impossible to detect at the data entry stage. Neither is it possible in the latter case simply to combine all the occurrences of d' with the following word, since it is sometimes completely absent.

70 For example, assuming that full names were consistently recorded in full on the cemetery lists, and taking all 5,418 males with at least two names for whom all component names are known, proportions with António or José occurring as one of the first two component names are calculated (with * denoting any name) as $P(\text{António}, *) = 0.15$, $P(*, \text{António}) = 0.06$, $P(\text{José}, *) = 0.19$, $P(*, \text{José}) = 0.12$, $P(\text{António}, \text{José}) = 0.02$, $P(\text{José}, \text{António}) = 0.02$. Thus, while $P(\text{António}, \text{José}) = 0.024$, $P(\text{António}, *) \bullet P(*, \text{José}) = 0.017$.

more component names. Practical considerations then led to the omission of one or more component names from manuscript documents – notably, the electoral registers.⁷¹ Today, in order to overcome some of these problems, Portuguese nationals are often asked to ‘state all names in the order in which they appear on the identity card or passport’;⁷² in the nineteenth century, however, this was unfortunately not so. As an illustration of some of the problems, several of the full names of individual-entries referring to Manuel Felix Mancio da Costa Barros (the mayor of Viana between 1855 and 1859) are shown below.

Manoel Felix Mancio da C^{ta}
 M^{el} Felix Mancio da Costa Barros
 Manoel Felix M^{cio} da Costa Barros
 Manoel Felix Mancio da C^{ta} Bar^{os}

Occupation data suffer all the aforementioned problems of textual data without the compensatory advantage of fairly consistent or obvious abbreviations.⁷³ In fact, particularly on the electoral registers, it appears that the scribes may have recorded occupations in extremely abbreviated form because they were familiar with many of the electors and would therefore know the full occupation of an individual upon identifying who he was.⁷⁴ This creates the feedback effect that while occupation might be used for record linkage, sometimes linkage is required in order to identify occupation accurately. Finally, in some cases the occupational description is perfectly legible but initially quite unfamiliar. Where these occupational descriptions are not later found in dictionaries or encyclopaedias,⁷⁵ they must be treated as unknown.

Birthplace information, although usually recorded without abbreviations, and sometimes supplemented by *termo de* (in the neighbourhood of [a larger town]), is often still quite ambiguous.⁷⁶ This variable also suffered from the use of *idem*. Finally, although it appears frequently on all but the electoral registers, its usefulness for record linkage is further limited because the majority of the population was born in Viana itself.

71 For example, the second name José was sometimes omitted, more often however, the last of two or more second names was omitted.

72 For example, European Communities Social Security Regulations – Certificate of Entitlement to Benefits in Kind During a Stay in a Member State [E111] [GB], Note 1a.

73 For example, the occupation *Am^{se} da R^{da} F* was recorded on several electoral registers before it appeared in full as *Amanuense da Repartição da Fazenda*.

74 For example, the occupation *P^{do} Lyceu* was used for both *Porteiro* and *Professor do Lyceu*.

75 For example, the occupation *lampianista* was eventually found in the *Grande Enciclopédia Portuguesa e Brasileira* in Porto's municipal library, a *lampianista* was a public gas lamp cleaner.

76 For example, the birthplace given as Braga could have been referring to the City, the Borough, or the District of Braga. In addition, sometimes the patron saint of the parish of birth was given, rather than the parish itself, so that the birthplace São Miguel could have been referring to any of four parishes in the Borough of Viana, the 15 in the rest of the district, or even one further afield.

Marital status is subject to two quite different forms of uncertainty. First, only a handful of people (usually females with children) were explicitly recorded as being single on the muster-rolls and in the passport books. However, the importance of marital status in indicating the availability of an individual for military service would suggest that the majority of those for whom no record was made are likely to have been single. Second, on the electoral registers and cemetery lists, the recording of this variable suffered very heavily from the use of *idem*.

Numerical data generate much less ambiguity of course, only being affected by illegibility and inaccuracy.⁷⁷ Dates of birth were sometimes only recorded to the month, or even the year, and were initially the cause of a certain amount of confusion on the muster-rolls, where they sometimes appear as a combination of characters and digits.⁷⁸ However, where available and accurate, date of birth provides an invaluable variable for record linkage. Of the sources considered here, it only appears on the muster-rolls, where it facilitates the identification of intra-urban migrants enormously.

Age, although fairly consistently recorded, was very inaccurate, on the early electoral registers, a discrepancy between the ages recorded for one individual in adjacent years of up to 10 or 12 years is not unusual. By the 1870's, however, the rolls were being updated from the previous year, reducing the problem with respect to the linkage of sequential registers enormously, though not necessarily providing greater absolute accuracy.

5.3 Coding and Standardisation of the Data

Marital status information, as mentioned in Section 5.1, was coded when the data were originally entered. The coding is designed in such a way as to be able to distinguish, for example, between individuals explicitly recorded as married and those whose recorded marital status, *idem*, referred back to some previous individual explicitly recorded as married.⁷⁹

Textual data were retrieved and listed alphabetically so that variations in abbreviation and spelling generally appeared close together, enabling lookup tables to be created manually. Each textual variable was then given a standardisation code which points to its standardised version, sometimes, where no standardised version already existed, one would be created. In addition, besides being standardised, text was coded for the purposes of record linkage and subsequent analysis.

Two occupational groupings were adopted. First, the twelve occupation function groups used for the Portuguese census of 1890 were incorporated.⁸⁰ Unfortunately, no records of which occupations belonged to which group exist, but

⁷⁷ Inaccuracy in numerical data is caused by transcription, transposition, and other errors, of these, Smythe (1968) finds that the most difficult to correct – transcription errors – are by far the most common (accounting for 80% of the errors in his data set).

⁷⁸ For example, the month written as *9 bro* referred to November not September.

⁷⁹ Marital Status code 0 – Unknown, 1 – Single, 2 – Married, 3 – Widowed, 4 – Clergy, 5 – Divorced, 6 – Single *idem*, 7 – Married *idem*, 8 – Widowed *idem*, 9 – Clergy *idem*.

⁸⁰ Occupation function code 0 – Unknown, 1 – Agriculture, 2 – Hunting & Fishing, 3 –

almost identical groupings for the 1834 census of the Iberian port of Gibraltar were referred to (Howes, 1950 142–157) In addition, a thirteenth occupation category – Ambiguous – was included Imprecise occupational descriptions which could be members of more than one group, such as ‘captain’ which could be a captain in the armed forces or a ship’s captain in the professions group, were placed in this category Second, contemporary grouping by economic sector was also used⁸¹ In both cases the code is slightly modified to include a code for those with no recorded occupation⁸²

A code for places was devised that could be used for both birthplace and destination (on the passports) It comprises 11 digits the first refers to the continent, the second to country, the next two to province, then two for district, two for borough, two for settlement (city, town, or village), and finally, one for parish The code therefore corresponds with administrative rather than geographical boundaries Two main types of ambiguity arose First, where several places with the same name existed, the place was coded as being the nearest one of that name Second, as is the case with Viana, a place name was often common to a district, borough, and city In coding the data therefore, the place ‘Viana’ was assumed to mean the Town/City, except where otherwise indicated, the names of other boroughs in the District of Viana were coded as the borough rather than its principal town, and outside the District of Viana, district names were coded as the district rather than their principal boroughs or cities In this way, as much ambiguity as possible is avoided, and inaccuracy is minimised

Finally, with respect to names, as mentioned in Section 4.2, each name is split up into its component names – a title (if any), a first name, and other, second names, these three types of component names are then standardised separately There are two important reasons for this First, abbreviated component names can be standardised according to their position within the full name⁸³ Second, while first names are important in the identification of the sex of individual-entries, the exact form of second names is less important Also, certain component names, which are sometimes difficult to differentiate between, are simply standardised together In order to illustrate these points, a

Mining, 4 – Trades & Industry, 5 – Transport, 6 – Commerce, 7 – Armed Forces, 8 – Public Administration, 9 – Professions, 10 – *Proprietários*, 11 – Domestic Service, 12 – Unproductive/Unclassified, 13 – Ambiguous The category *Proprietário* is somewhat problematic, it was introduced during the second quarter of the nineteenth century, and literally meant someone whose livelihood was derived from the rent collected on property which he owned However, the status it implicitly carried ensured its growth in popularity beyond its original meaning Thus, although the 1890 census enumerators were trained to ask probing questions in order to classify members of the population correctly (Recenseamento da População, 1890 XX-XXI), those drawing up other documents were less careful, thus generating pseudo *Proprietários* on these other documents Other manuscript sources from Viana suggest that some of these pseudo *Proprietários* were engaged in agriculture

81 Occupation sector code 0 – Unknown, 1 – Primary, 2 – Secondary, 3 – Tertiary, 4 – Unproductive, 5 – Unclassified

82 See Armstrong (1972) for a discussion of the coding of occupational information

83 For example, Lour^o is standardised to Lourenço if it appears as the first component name, Loureiro otherwise

Table 3. Example Component Name Standardisations

Type	Prefix	Raw Component Name	Expansion of Abbreviation	Standardisation
0		P.e	Padre	
1		Albano		Alb-no
1		Albino		Alb-no
1		Alvano		Alb-no
1		Alvino		Alb-no
1		Ellario		Ilario
1		Hylario		Ilario
1		Ilario		Ilario
1		Lour o	Lourenço	Lourenco
2		Lour o	Loureiro	Loureiro
2		Agonia		Agonia
2	da	Gonia		Agonia
2		Sa		Sá/Silva
2		S a	Silva	Sá/Silva
2		Cerq a	Serqueira	Serqueir-
2		Cerqueiras		Serqueir-
2		Cerqueiro		Serqueir-
2		Segueira		Serqueir-
2		Serqueira		Serqueir-
2		Feixeira		Teixeira
2		Teixeira		Teixeira

number of standardisations are shown in Table 3

In conclusion, the standardisation process allows for the individual standardisation of variables to the extent required for record linkage. Although this might initially appear to be a somewhat *ad hoc* procedure, it is potentially more powerful than automatic techniques in common use (e.g. Soundex or Viewex). In short, any and all the coding and standardisation problems outlined in this Section can be overcome using this system.

6. RECORD LINKAGE OF THE VIANA DATA

This Section describes the record linkage of the Viana data in detail. First,

record linkage within the muster-rolls is presented. This precedes the more detailed discussion because, to a great extent, the muster-rolls explicitly provide their own links. Indeed, the record linkage of the muster-rolls enabled many of the problems of record linkage of the Viana data to be assessed. Second, a truly linked example individual life history is presented, this introduces the discriminatory power of the standardised and coded data, and provides an illustration of typical migration patterns among the urban elite. Next, the discriminatory power of the variables available for record linkage is discussed more generally. Finally, the linkage variables, criteria, and procedures adopted for the automatic record linkage within and between the different sources comprising the Viana data are described in detail.

6.1 Linkage Within the Muster-Rolls

To a great extent, the muster-rolls explicitly provide links between the different entries of persons who appear more than once. Therefore, the record linkage of the muster-rolls enables many of the problems of record linkage of the Viana data (e.g. the recording of Portuguese names) to be assessed.

Explicit links between entries are available because, where persons moved from one household to another, their movements are usually recorded on the rolls as observations such as *Passou para Rua Nova de Santa Anna N^o 7*⁸⁴. Further, the muster-rolls include a large number of persons – mostly adult females and young males – who share the same full name (because few component names were recorded), and for whom little other information is recorded. Under these circumstances, the record linkage within the muster-rolls was initially performed using only explicitly recorded links. Later, this linkage was scrutinised and supplemented using other fields with significant discriminatory power.

Of the 6,858 muster-roll entries, 501 have an intra-urban movement recorded as an observation, and a name field⁸⁵. The record linkage of these entries using their explicitly recorded links was performed in two stages, each stage was automatic⁸⁶ except that where a muster-roll entry was to be inserted in an existing chain of entries, or a choice of links existed, the program presented the fields of each record and requested the NIDs of the new chain, of course, this too could have been automated, but the number of cases involved did not justify the amount of programming that would have been required. First, links were created where one and only one individual with a consistent full name appeared in the

84 On the one hand, observations sometimes only refer to another road of Viana, without specifying any particular household, on the other hand, in cases where movement was to another household on the same road, only the number of the household is recorded.

85 A further 17 entries – infants – have a recorded intra-urban movement but no available name, these can be linked using the links of other members of the same household.

86 The program processes the entire database, identifying cases for which an observation record begins with the character string '*Passou para rua*', the immediately following road and household number are then used to search the household referred to (this is possible because the muster-roll NIDs are initially generated from the road and household number). Upon completion of this search, the program resumes processing the entire database.

household or road referred to, this resulted in 357 linkages (17 interactive) – no match could be found for 144. Second, possible links where the first standardised component name matched, and one set of component names was a subset of the other were examined interactively. This resulted in a further 66 linkages (28 interactive), 13 possible linkages were rejected because of inconsistencies in other fields, and, no possible matches could be found at all for the remaining 65.

The possible existence of other links within the muster-rolls – caused by the failure of the authorities to record all intra-urban movements – is most easily investigated using the date of birth field. With a population of certainly less than 10,000 during the period 1826–1833, the likelihood of two male residents of Viana sharing the same date of birth and the same first component name is low⁸⁷, it is therefore reasonable to create links between entries for which these fields are consistent. This approach created 133 new linkages, of which 31 were added to previously existing chains⁸⁸. Further interactive exploration using looser date of birth and name matches provided another 51 linkages.

In conclusion, it can be seen that while the explicit links recorded on the muster-rolls generated 436 linkages, a further 184 could be added using date of birth and name matches. There is little doubt that some links have been missed using these strategies, and that some of the 620 linkages are spurious, but these errors are minimised, certainly there is no evidence to suggest otherwise.

The muster-roll links described in this Section are used for the analysis of population and migration presented by Kitts (1988 Chapter 5), but they have subsequently been overwritten by links generated using the fully automatic procedures described in Sections 6.4 and 6.5.

6.2 An Example Individual Life History

A truly linked example individual life history is presented in Table 4. This serves two purposes. First, it introduces the discriminatory power of the standardised and coded data. Second, it provides an illustration of typical migration patterns among the urban elite.

João da Silva São Miguel probably moved to the Town of Viana in about 1828, because he first appears on the muster-rolls in the second ink and handwriting of the First Company. Soon thereafter, he moved from his first place of residence – a household in which a family of merchants lived – to an apartment of his own, less than 100 metres away. Later, he married in 1836, and remained based in Viana until he died in 1878, he appears on every electoral register between

⁸⁷ This point is illustrated by considering a hypothetical, closed, stable population in which every person lives for exactly 35 years – about 12,783 days. With a population of around 4,000 persons, the probability of a person being born on a particular day is 0.31 (4,000/12,783), the probability of two persons being born on that day is 0.10 (0.31²). Then, with, say, 35% of the population sharing the most common first component name, the probability of two persons being born on one day and being given the same first component name is 0.03 (0.35 • 0.31²).

⁸⁸ In fact, 138 possible linkages were generated, but 5 were rejected because of inconsistencies in other fields.

Table 4. The Life History of João da Silva São Miguel

Muster-Rolls

Date of Birth 21/12/1808
Birthplace São Miguel de Carreira (Braga District)
Occupation Cloth Merchant
Address 1, Rua do Caes, 1, Rua do Villarinho

Electoral Registers 1834, 1835, , 1878

Birthplace São Miguel de Aves (Porto District)
 São Miguel (Viana District)
 Matriz *idem* (Viana District)
 São Miguel de Setelada (Unknown)

Age discrepancy From 42 Electoral Rolls
 Minimum 0, Maximum 9, Mode 2,
 Mean - 0 07, Standard deviation 3 23

Marital Status Single -1836, Married 1840-
Occupation Merchant

Cemetery Lists

Date of Death 29/07/1878
Birthplace São Miguel de Junteira (Unknown)
Marital Status Married
Address Rua do Caes

Occurrences of His Sons

João da Silva São Miguel Junior was issued with a passport for Brazil in 1859 (aged 21). He does not appear on any other documents, so it is unlikely that he ultimately returned to Viana.

Manuel da Silva São Miguel appears on the electoral registers between 1880 (aged 41, single, merchant) and 1911. He married between 1883 and 1888, but his wife died between 1891 and 1894. He appears on the cemetery lists in 1919.

António da Silva São Miguel was issued with a passport for Rio de Janeiro in 1856 (aged 14). He returned before 1880, when he began to appear on the electoral registers (single, merchant), he appears on the cemetery lists in 1910.

José da Silva São Miguel was issued with a passport for Brazil in 1861 (aged 18, occupation *caixeiro*). He does not appear on any other documents, so it is unlikely that he ultimately returned to Viana.

1834 and 1878, and on the cemetery list of the latter year. He and his wife had at least four children, of which three were sent to Brazil, thereby avoiding military conscription. His wife died in 1906.

The linkage of records referring to João da Silva São Miguel is particularly straightforward because his name suffered little variation, it was always recorded in full, including all his component names. Also, apart from one of his sons, nobody else with the same full name appears to have existed in Viana during the period of study.⁸⁹ Further, his wife and children are also easily identified, since no other family appears to have had full names terminating with the components São Miguel. Thus, with respect to names, the Silva São MIGUELS are certainly not typical. At the same time, however, the record linkage of the São MIGUELS is almost certainly correct, so that variations in other items of information concerning João da Silva São Miguel can be used to provide an introductory illustration of the discriminatory power of these fields.

First, age discrepancies can be investigated under the assumption that João da Silva São Miguel's date of birth was accurately recorded (twice) on the muster-rolls. As can be seen, there was sometimes a difference of up to 9 years between his actual and recorded age. Second, it is known from the marriage records of Santa Maria Maior that João da Silva São Miguel (birthplace São Miguel de Carreira, Borough of Barcellos, District of Braga) married Rosa Maria d'Amorim (birthplace Santa Maria Maior) on 10 February 1836, therefore, his recorded marital status is always accurate. Third, occupational information on João da Silva São Miguel is also always accurately recorded, and always consistently coded – as commerce in the occupation function code, and tertiary in the occupation sector code. Finally, considering birthplace, it is likely that João da Silva São Miguel personally provided his birthplace for his muster-roll entries and marriage record, while the authorities drawing up the electoral registers did not consult him directly, and those drawing up the cemetery lists simply could not. This may explain why the five birthplaces given on the latter two documents differ from each other, and from the three birthplaces recorded elsewhere. As can be seen, his first recorded birthplace on the electoral registers is erroneous, his second (simply São Miguel) is wrongly coded as the nearest parish with that patron saint, his third ([Matriz] *idem*) is also erroneous, and his fourth, together with that recorded on the cemetery list, are unidentified.

6.3 Discriminatory Power of Record Linkage Variables

In this Section, the discriminatory power of the variables available for record linkage is discussed more generally. In order to explore the issues raised, the

⁸⁹ The slight complication that arises because one of João da Silva São Miguel's sons is also called João da Silva São Miguel is easily resolved using age information.

records of males appearing on the cemetery lists in the period 1875–1899 are examined closely. Attention is restricted to these 1,844 records⁹⁰ for two reasons. First, young children, who only appear on the earlier cemetery lists, and whose first component name alone was usually recorded, are excluded. Second, the extent of inter-generational repetition of full names is reduced. Of course, the advantage of this analysis is that individuals rarely appear more than once on the cemetery lists,⁹¹ so that potential links are known to be false.

Names are without doubt the most powerful variable consistently available for record linkage. However, while the standardisation and coding of component names overcome the problems of variation in the recording of component names, they fail to resolve those of component name omission or reordering. These problems can be approached using search algorithms which explore the database for occurrences of component name subsets, generating files of potential links. In essence then, the standardised full name variable can be used to bring together all possible links. The resulting chains can then be divided using other information.

To explore these issues, the frequencies with which the standardised full names of the example records appear more than once are shown for different numbers of component names in Table 5. (It is to be noted that this analysis is performed under the assumption that names were consistently recorded in full, which may not have been the case.)

Table 5. Full Name Repetition on the Cemetery Lists

Frequencies	Number of Component Names					
	1	2	(%)	3	(%)	≥4
1	3	317	81	911	94	280
2	0	46	12	47	5	3
3	0	14	4	5	1	0
≥4	0	13	3	2	0	0
Total (Different Names)	3	390	100	965	100	283
Total (Occurrences)	3	527		1028		286

The figures indicate that by linking on standardised full name only, less than 20% of links are false. Further, by considering the sub-population with full names consisting of three or more component names only, this figure reduces to about 5%. Where linkage of the male urban elite is concerned, the figure, though difficult to quantify, ought to be even smaller.

⁹⁰ The sample excludes a few individuals for whom one or more component names are not known, for example, that of a body washed up on Viana's beach.

⁹¹ A small number of individuals do appear on the cemetery lists more than once, but the fact that their corpses were dug up and then reburied is usually recorded.

Also, the 1,248 different standardised full names consisting of three or more component names were examined in order to find how many of them occur with an identical first component name as subsets of one or more of the 283 names consisting of four or more component names, 84 do, in a total of 104 longer names. Of these matches, 73 involved the inclusion or addition of one more component name, in 13 cases between the first and second components, in 17 cases between the second and third components, in 40 cases at the end of a three-component name, and in the other 3 cases at the end of a four-component name. All the remaining 11 matches involved the inclusion or addition of two more component names.

Next, age information is considered. Age is the most powerful variable for differentiation between different persons with identical standardised full names. For example, age information is available for 42 of the 104 false matches identified above, the age discrepancies associated with these 42 pairs of records⁹² have the following statistics: minimum 2, maximum 76, mean 26.1, and standard deviation 19.7. With respect to variations in the recorded ages of individuals, the truly linked example of the previous Section provides a useful illustration. Statistics of the 42 available age discrepancies are: minimum 0, maximum 9, mode 2,⁹³ mean -0.1, and standard deviation 3.2. Further, the number of individual ages recorded as being greater than 90 on the electoral registers suggests that age discrepancy increases with actual age.⁹⁴ Nevertheless, the evaluation of the discrepancy between two ages is not straightforward. Having calculated an approximate year of birth from available date of birth or age information, two approaches might be adopted. First, a fixed cutoff point for the discrepancy is considered. Where fathers and sons share identical standardised full names, a difference of 15 years would usually provide the correct division of records, however, where persons not directly related share identical standardised full names, a 15 year cutoff point for age differences would sometimes fail to provide adequate division.⁹⁵ Second, the discrepancy might be scored, this approach is adopted in the linkage of the Viana data, and is described in Section 6.4.

Finally, marital status, occupation and birthplace information are considered. On the one hand, marital status and occupation information are rarely as accurately recorded as for the truly linked example of the previous Section. On the other hand, birthplace information rarely presents the complete lack of accuracy evident in the truly linked example of the previous Section, nevertheless, it is scarcely informative for record linkage since it is only recorded on a handful of electoral registers, and because a large proportion of the population was born in

⁹² These age discrepancies are 2, 2, 3, 5, 5, 5, 6, 7, 8, 10, 11, 12, 13, 14, 14, 15, 16, 20, 20, 20, 21, 21, 22, 23, 26, 27, 28, 29, 32, 38, 40, 41, 41, 41, 42, 44, 47, 51, 57, 67, 75, and 76.

⁹³ The modal age discrepancy of 2 years arises because, in the example given, age was for some years updated from the previous year's electoral register.

⁹⁴ Some researchers, however (e.g. Nitzberg, 1968), find no evidence of any correlation between age and magnitude of age discrepancy.

⁹⁵ For example, 16 of the 42 age discrepancies reported above were less than or equal to 15.

the Viana itself

In conclusion, the Viana data can certainly not be described as being of high quality. However, considering the two main problems of record linkage (variation and errors in identifying items, and duplication of identifying item sets), the techniques described in previous sections allow the former problem to be reduced considerably, and, because the male urban elite are of primary interest, the latter problem is relatively small.

6.4 Linkage Variables and Criteria

In this Section, the linkage variables and criteria discussed generally in the previous Section are described explicitly. First, the variables extracted from each record are described, then the match scoring functions developed to measure the consistency of pairs of these variables are described. Finally, the application of these variable match scoring functions using a record match scoring function is described.

The most important variable is the NID, from this variable, three linkage variables are extracted: document (DOC), parish (PAR), and year of entry on document (YOE). The next most important variable is the standardised full name (SFN), this is extracted as a series of standardised component name codes (SCNCs) together with their respective frequencies of occurrence in the data base (CNFs), and the number of component names (NOCN) is separately recorded. Other variables extracted are: birthplace code (BPC), estimated year of birth (EYB), available directly from date of birth or indirectly from age, marital status code (MSC), occupation function code (OFC), and occupation sector code (OSC).

The match scoring functions developed to measure the consistency of these variables essentially return one of several values, they return a value of unity when insufficient information is available, a value greater than unity when the matching criteria are met, and a value less than unity when they are not. The match scoring functions are: name (FNAM), document (FDOC), year of entry (FYOE), parish (FPOE), birthplace (FBPC), birth and residence interaction (FBOR), year of birth (FEYB), age at entry (FAAE), marital status (FMSC), occupation (FOCC), duplicate electoral (FDEL), and record (FREC).

The name match function compares the standardised full names of two records. The two standardised full names are considered to be consistent if they share the same first component name, and all the component names of the shorter name occur in the longer name. If the two standardised full names are inconsistent, the name match score (FNAM) is set to 0, otherwise, it is set according to a comparison of the numbers of component names: if both standardised full names have more than four components, it is set to 2, if both have more than five, it is set to 3, if both have more than six, it is set to 4, and, if both have seven, it is set to 5.

The document match function checks two criteria: that individuals only appear once on the electoral registers of any year, and that individuals do not reappear after having been buried. If these criteria are met, the document match score

(FDOC) is set to 1, otherwise it is set to 0

The year of entry match function is incorporated in order that records with no age, marital status, occupation, or birthplace information can be linked to records with close years of entry. If the absolute value of the difference between the years of entry (AYE) is less than 15, the year of entry match score (FYOE) is set to just greater than 1

$$\text{FYOE} = 1 + (15 - \text{AYE}) \bullet (0.0006)$$

Further, an explicit allowance is made for records from the electoral registers drawn up in the years 1837–1839 (in which only names were recorded). If one of the records is from an electoral register drawn up in the years 1837–1839, and the other record is from an electoral register drawn up for the same parish in the years 1834–1842, the year of entry match score is set to just greater than 2

$$\text{FYOE} = 2 + (7 - \text{AYE}) \bullet (0.1)$$

The parish match function requires that both records refer to residents of the same parish, or that such information is unavailable. If these criteria are met, the parish match score (FPOE) is set to 1, otherwise it is set to 1/8

The birthplace match function checks that two places of birth are consistent. For example, if the parishes of birth are available for both records, then they must be identical, but where, say, only one parish and one borough of birth are available, the parish must merely be in that borough. If these criteria are met for a birthplace outside the Town/City of Viana, the birthplace match score (FBPC) is set to 6, if they are met for a specified parish of Viana, it is set to 3, if they are met for Viana with parish unspecified, it is set to 2, if information is unavailable it is set to 1, and if the criteria are not met, it is set to 1/2

An additional birth or residence interaction match function is included so that where place of birth and parish of residence match scores are unavailable but the place of birth of one record is the place of residence of the other (i.e. one of the two parishes of the Town/City of Viana), the birth or residence match score (FBOR) is set to 2, otherwise it is set to 1

The year of birth match function compares year of birth information. If one or both the estimated years of birth are unavailable, the year of birth match score (FEYB) is set to 1. Otherwise, the year of birth match score is set according to the absolute value of the discrepancy between the estimated years of birth. Further, because the recording of age was sometimes quite inaccurate on the electoral registers, the score allocated to the discrepancy depends on the combination of documents from which the records originate. Where records from the muster-rolls, passport books, or cemetery lists are concerned, the year of birth match scores are set as follows: for discrepancies from 0 to 4 years, the scores are 6.5, 6.4, 6.3, 6.2, and 6.1, for discrepancies from 5 to 14 years, the scores are 1.10, 1.09, 1.08, 1.07, 1.06, 1.05, 1.04, 1.03, 1.02, and 1.01, for discrepancies from 15 to 24 years, the scores are 1, 1/2, 1/4, 1/8, 1/16, 1/32, 1/64, 1/128, 1/256, and 1/512, and for discrepancies of more than 24 years, the score is set to 0. Where both records are from the electoral registers, the year of birth match scores are set as follows: for discrepancies from 0 to 4 years, the scores are 6.5, 6.4, 6.3, 6.2, and 6.1, for discrepancies from 5 to 14 years, the

scores are 4 6, 4 5, 4 5, 4 3, 4 4, 4 5, 4 4, 4 3, 4 2, and 4 1, for discrepancies from 15 to 24 years, the scores are 4 6, 4 5, 4 5, 4 3, 4 4, 4 5, 4 4, 4 3, 4 2, and 4.1, and for discrepancies of more than 24 years, the score is set to 1/24

An additional age at entry interaction match function is included to ensure that the age at entry implied by the linkage of year of birth and year of entry information from two records is reasonable. If the year of entry for one record precedes the year of birth implied by the other, then the age at entry match score (FAAE) is set to 0. If the linkage of a record from the electoral registers would imply that the age of the elector was less than 15, then the age at entry match score is set to 0. And, if the linkage of a record from the cemetery lists would imply an age at death of less than 0 or greater than 124, then the age at entry match score is set to 0. Otherwise, it is set to 1.

The marital status match function can be applied at two levels. The first, a strong marital status match, requires that the implied marital statuses are identical. The second, a weak marital status match, merely requires that an individual cannot have been explicitly recorded as single after they had been explicitly recorded as married, or widowed. Where a muster-roll record is involved, a strong marital status match is required, otherwise only a weak marital status match is required. These differences are incorporated because the muster-rolls were compiled during a relatively short period in which the variation of information may be assumed to be minimal, but the other documents were compiled over long periods in which recording is likely to suffer increased variation, and marital status will also have changed naturally. For a consistent marital status match (weak or strong, as appropriate), the marital status match score (FMSC) is set to 1, and for a poor match, it is set to 1/2. Also, irrespective of the documents from which the records originate, if the marital status of one record indicates that he is a member of the clergy, while the other record's marital status is specified as single, married, or widowed, then the marital status match score is set to 1/4.

The occupation match function requires that occupation function codes and occupation sector codes are identical, if they are, the occupation match score (FOCC) is set to 6. Otherwise, if the occupation function codes are identical, but the occupation sector codes are not, or the occupation function codes are not identical, but the occupation sector codes are, then the occupation match score is set to 1/2. Finally, if the occupation function codes are not identical, and the occupation sector codes are very different (i.e. a primary sector and a tertiary sector), then the occupation match score is set to 1/8.

Finally, an additional duplicate electoral match function is included. If both records are from electoral registers drawn up in the same year, then the duplicate electoral match score (FDEL) is set to 1/4, otherwise it is set to 1. This allows for two possibilities. First, certain electors are known to have sometimes appeared on the registers of both parishes drawn up in a particular year. Second, sometimes two (sometimes even three) electors with identical names and virtually indistinguishable other characteristics appear on the electoral register of one parish. This problem is discussed further in Section 6.7, which covers the performance, and limitations, of the record linkage.

The record match function evaluates each of the above match scores, and

calculates the record match score (FREC) as their product

$$\text{FREC} = \text{FNAM} \bullet \text{FDOC} \bullet \text{FYOE} \bullet \text{FPOE} \bullet \text{FBPC} \bullet \text{FBOR} \bullet \text{FEYB} \bullet \text{FAAE} \bullet \text{FMSC} \bullet \text{FOCC} \bullet \text{FDEL}$$

The record match function was tuned by adjusting the match scoring functions to emulate a human decision process. This was accomplished using a number of linkage examples from the Viana Database, including several of the most difficult. Also, all possible combinations of variable match scores were generated and analysed, to ensure that the record match scoring function was in some sense robust.

The tuning of the record match scoring function is best illustrated with a few examples. First, record match scores must be intuitively meaningful with respect to the variable match scores from which they are generated, for example, the combination of a good age match score and a poor occupation match score (due to occupational mobility, perhaps) must generate a higher record match score than that generated by a poor age match combined with a good occupation match. Second, peculiarities of the data being linked must be accommodated, for example, the existence of records with only age information, and the likelihood that certain electors sometimes appeared on the electoral registers of both parishes in a particular year. Third, important interactions between linkage variables must be examined, for example, the implied age at death resulting from the linkage of a record from the cemetery lists to a record with age information on another document.

In general, a record match score less than or equal to unity is considered as a poor record match, while a record match score greater than unity is considered as a good record match. The match scoring functions are summarised in Table 6.

Table 6. Match Scoring Functions

Function	Poor match	Unavailable	Good match
FNAM	0		1 to 6
FDOC	0		1
FYOE	1		1 0006 to 1 009
FPOE	1/8	1	1
FBPC	1/2	1	2, 3 or 6
FBOR	1	1	2
FMSC	1/4 or 1/2	1	1
FOCC	1/8 or 1/2	1	6
FEYB	0 to 1	1	1 to 6.5
FAAE	0	1	1
FDEL	1/4	1	1

The power of the record match function can be illustrated by completing the discussion of the micro-analysis presented in the previous Section. The 104

pairs of records identified in Section 6.3, for which the shorter full name occurs with an identical first component name as subsets of the longer name, were linked using all but one of the variable match scoring functions described – the document match scoring function was suppressed. 20 of the 104 pairs were linked, but it is important to emphasise that not one would have been linked if the document match scoring function had been included in the record match scoring function. These 20 pairs were scrutinised in order to ascertain the cause of the failure of the algorithm to identify their potential links as false. An example illustrates the problem clearly. António Joaquim Pereira was buried in 1889, he was born in Monserrate, and died aged 54, married, with no recorded occupation, António Joaquim Pereira Braga was buried in 1895, he was also born in Monserrate, he died aged 55, married, and was recorded as being a shoemaker. Unfortunately, in the record linkage of the Viana data, using the methodology and techniques adopted, it would be impossible not to link a pair of records with such similar characteristics, without failing to link other pairs of records which do pertain to the same historical person.

6.5 Linkage Strategy

The approach adopted essentially comprises five stages, incorporating a combination of agglomerative and divisive procedures. In Stage 1, all records with identical standardised full names are linked together into chains. In Stage 2, a search algorithm identifies pairs of chains for which standardised full names indicate a possible match – a possible match is defined as occurring when two standardised full names share the same first component name, and all other component names of the shorter name occur in the longer name. These chains are linked together to form larger chains. In Stage 3, the chains of records are divided into blocks of records, blocks of records are defined as having the same number of component names, and identical sets of component names, with the same first component name, but other component names not necessarily always in the same order. In Stage 4, the blocks of records are further divided into sub-blocks using supplementary information, these sub-blocks represent different historical persons with identical sets of standardised component names. Finally, in Stage 5, each chain of blocks and sub-blocks is dealt with separately. First, links between sub-blocks of different blocks are generated, starting with the most likely name matches. Then, again drawing upon supplementary information, a set consisting of the records of one historical person is extracted.

Thus, Stages 1 & 2 are agglomerative, they ensure that all possible matches are linked together within chains. Stages 3 & 4 are divisive, they break down the chains into blocks and sub-blocks, according to the name variable alone, and then using supplementary information. Finally, Stage 5 is again agglomerative, in that sub-blocks of different blocks are collected, enabling a set consisting of the records of one historical person to be extracted. In the remainder of this Section, the five stages comprising the record linkage strategy are described in more detail.

In Stage 1, all records with identical standardised full names are linked together into chains. Stage 1 linkage statistics are shown in Tables 7 and 8, for

males and females respectively. Considering males, it can be seen that while the mean number of records per set is greater than four for names with four or less component names, it is about two for names with five or more component names. This reflects two phenomena: first, the likelihood of two individuals sharing a full standardised name decreases as the number of component names increases, second, individuals with many component names were often recorded using only a subset of those names.

Following Stage 1, only records belonging to males with full names consisting of three or more components are considered. This is in order that attention is focused on records for which the name variable has substantial discriminatory power. In the context of this research, this does not present a problem because the male urban elite are of primary interest. However, it is noted that it is perfectly possible for a name with two components to be rarer than another name with three or more components, so the exclusion of names with two or less components represents a somewhat simplistic approach.

In Stage 2, a search algorithm identifies pairs of chains for which standardised full names indicate a possible match – a possible match is defined as occurring when two standardised full names share the same first component name, and all other component names of the shorter name occur in the longer name. These chains are linked together to form larger chains. According to these criteria, any Stage 1 chain may be linked to many other Stage 1 chains, so the resultant Stage 2 chains often comprise a number (sometimes a very large number) of Stage 1 chains. The effect of ignoring component name order for all but first component names is illustrated in Table 9. A comparison of Tables 7 and 9 indicates that for names of three to six components, the number of sets has decreased. For example, considering names with six components, it can be seen that two names occurring just once each are identical except for the order of their third to sixth component names, thus, in Table 9, the number of sets of just one record has fallen by two, while the number of sets of exactly two records has risen by one.

Following Stage 2, only chains of more than one record are considered, this is because chains of only one record represent sets consisting of the records of only one historical person.

In Stage 3, the chains of records are divided into blocks of records, these blocks of records are defined as having identical sets of standardised component names, but not necessarily always in the same order. The blocks correspond to the sets of records shown in Table 9.

In Stage 4, the blocks of records with identical sets of standardised component names are divided into sub-blocks using supplementary information, these sub-blocks represent different historical persons with identical sets of standardised component names. Sub-blocking is performed as follows. First, the records of a block are sorted by NID. Thus, the records from each document type are processed separately, muster-roll records are considered first, then electoral register records, then passport book records, and finally cemetery list records, and, within the latter three document types, the records are sorted in time. Then, each document type is processed separately, and within each document type, each year of entry is processed separately. Each record being processed is compared with each existing sub-block of the block, and, if appropriate, linked

to it. If none of the comparisons with existing sub-blocks are satisfactory, then a new sub-block is created to contain the record. For each comparison, the record is compared with every record in a particular sub-block, and the geometric mean of all the record match scores is calculated. If this average record match score exceeds unity, the sub-block is linked to the sub-block. Where several average record match scores exceed unity, the record is inserted in the sub-block with the highest average record match score.

In certain circumstances, the sub-blocking process therefore becomes slightly complicated, for example, where more than one electoral register record exists for the same year of entry within a single block (i.e. two or more electors in a particular year had matching names). Consider the sub-blocking of the block with standardised full name Manuel José Rodrigues. By the time the sub-blocking process has reached the year of entry 1854 on the electoral registers, two sub-blocks have been created. In 1854, four electors were recorded with the standardised full name Manuel José Rodrigues, two in the parish of Santa Maria Maior, and two in Monserrate. The sub-blocking process calculates average record match scores for the comparison of each of these four electoral register records with each of the two existing sub-blocks, thus, eight comparisons are made. The comparison record and sub-block with the highest average record match score are then linked, and the sub-blocking process recalculates the comparisons of the three remaining records with the sub-block in which the first record has been inserted. This iterative process continues until all the four electoral register records have been allocated to sub-blocks.

Linkage statistics summarising the effects of Stages 2, 3 & 4 are shown in Table 10. It can be seen that the 39,079 records comprise a total of 8,725 blocks, these blocks have been allocated to 6,531 chains, and have been divided into 11,103 sub-blocks. Thus, at the end of Stage 4, a minimum of 6,531, and a maximum of 11,103 historical persons have been identified.

æ Following Stage 4, only chains of more than one sub-block are considered, this is because chains of only one sub-block represent sets consisting of the records of only one historical person.

Finally, in Stage 5, each chain of blocks and sub-blocks is dealt with separately. First, links between sub-blocks of different blocks are generated, starting with the most likely name matches. Then, again drawing upon supplementary information, a set consisting of the records of one historical person can be extracted. Two distinct procedures are performed on each chain.

In Stage 5A, the block of records whose standardised full name comprises the largest number of component names – the master block – is selected from the chain. Then, any other blocks whose standardised full names comprise a subset of these master component names are selected from the chain. All the records selected are stored in a sub-chain. Next, the standardised component names of the master block are notionally reordered by frequency of occurrence, and their codes are stored in a variable. Then, the standardised component names of all the other blocks in the sub-chain are similarly reordered. Finally, the blocks of the sub-chain are sorted by this notionally-reordered-standardised-component-name-code variable. This procedure results in an ordering of blocks by some measure of the likelihood that members of the block match members of the

Table 10. Stages 2, 3, & 4: Record Linkage Statistics

Component Names	Blocks in Chain										Chains Blocks Sub-blocks Records			
	1	2	3	4	5	6-10	11-50	51-500	501-1000					
One-block Chains														
3	4814	-	-	-	-	-	-	-	-	4814	4814	5612	17991	
4	985	-	-	-	-	-	-	-	-	985	985	1005	2489	
5	121	-	-	-	-	-	-	-	-	121	121	124	199	
6	19	-	-	-	-	-	-	-	-	19	19	19	28	
7	2	-	-	-	-	-	-	-	-	2	2	2	2	
Multi-block Chains														
3&4	-	333	101	35	6	14	2	-	-	491	1276	1749	7864	
3&5	-	21	2	-	-	-	-	-	-	23	48	63	287	
3&6	-	-	-	2	-	-	-	-	-	2	6	7	8	
3&7	-	1	-	-	-	-	-	-	-	1	2	2	4	
3,4&5	-	-	7	8	9	7	2	3	-	36	646	1081	4406	
3,4&6	-	-	2	-	-	1	-	-	-	3	12	13	54	
3,4&7	-	-	-	1	-	-	-	-	-	1	4	5	7	
3,5&6	-	-	1	-	-	-	-	-	-	1	3	4	5	
3,6&7	-	-	1	-	-	-	-	-	-	1	3	3	17	
3,4,5&6	-	-	-	-	2	3	-	-	-	5	35	47	129	
3,4,5,6,&7	-	-	-	-	-	-	-	2	-	2	700	1312	5296	
4&5	-	22	1	-	-	-	-	-	-	23	47	53	291	
4&6	-	1	-	-	-	-	-	-	-	1	2	2	2	
	TOTALS										6531	8725	11103	39079

Note that there is a slight discrepancy between Tables 9 and 10. On the one hand, in Table 9 there are a total of 10,960 Blocks and 50,271 records, or, considering records with three or more component names only, 8,768 Blocks and 39,155 Records. On the other hand, in Table 10 there are a total of 8,725 Blocks and 39,079 Records. The differences arise because the statistics for the two tables were calculated at stages of database development between which the raw data were edited, and a number of records with erroneous names were deleted.

master block ⁹⁶

In Stage 5B, each sub-block of each block in the sub-chain is compared with the master block, and, if appropriate, linked to it. For each comparison, every record in the sub-block is compared with every record in the master block, and the geometric mean of all the record match scores is calculated; thus, if there

⁹⁶ For example, a block with several relatively rare component names in common with the master block and one relatively common component name missing, is more likely to contain records which pertain to the same historical person as those of the master block than a block with several relatively common component names in common with the master block and one relatively rare component name missing.

Table 11. Final Record Linkage Statistics

Component Names	Blocks in Chain						Chains	Blocks	Records
	1	2	3	4	5	6-10			
One-block Chains									
3	7099	-	-	-	-	-	7099	7099	25489
4	1728	-	-	-	-	-	1728	1728	5690
5	198	-	-	-	-	-	198	198	402
6	34	-	-	-	-	-	7	7	10
7	7	-	-	-	-	-	7	7	10
Multi-block Chains									
3&4	-	467	67	11	1	-	546	1184	6025
3&5	-	35	4	2	-	-	41	90	367
3&6	-	4	2	-	-	-	41	90	367
3&7	-	-	-	-	-	-	-	-	-
3,4&5	-	-	12	6	1	-	19	65	331
3,4&6	-	-	1	-	-	-	1	3	14
3,4&7	-	-	-	-	-	-	-	-	-
3,5&6	-	-	1	-	-	-	1	3	23
3,6&7	-	-	-	1	-	-	1	4	9
3,4,5&6	-	-	-	-	5	10	2	13	66
3,4,5,6,&7	-	-	-	-	-	-	-	-	-
4&5	-	37	6	-	-	-	43	92	476
4&6	-	2	-	-	-	-	2	4	4
	TOTALS						9731	10544	39079

are r_m records in the master-block, and r_s records in the sub-block, the average score will be the geometric mean of $r_m \bullet r_s$ record match scores. If this average record match score exceeds unity, the sub-block is linked to the master-block. Having then extracted all the records of one historical person, and returned any remaining records of the sub-chain to the main chain, the linkage program returns to Stage 5A.

Final linkage statistics are shown in Table 11. It can be seen that from the 39,079 records, 9,731 historical persons have been identified, this figure lies between the minimum (6,531) and maximum (11,103) number of historical persons identified at the end of Stage 4. Of these, 9,057 individual chains comprise records with identical standardised full names, however, a total of 666 individual chains (7% of all individual chains) have only been created because the linkage strategy specifically allows for component name omission and reversal. Indeed, one particular individual chain comprises eight different blocks. Gonalo Joaquin Almeida Sousa S Baptista appears on the muster-rolls once, on 26 electoral registers, and on the cemetery lists once. There is virtually no doubt that all these 28 records pertain to a single historical person,

since there is a high degree of agreement among the variables, his year of birth is always between 1780 and 1785, he was always employed in the customhouse, and, presumably, he got married between 1836 and 1842. Examples of the eight different combinations of his component names are

NID	Standardised Full Name
1-17-003-01	Gonçalo Joaquin Almeida Sousa Sá Baptista
2-854-2-058	Gonçalo Joaquin Almeida Sousa Baptista
2-835-2-029	Gonçalo Joaquin Almeida Sá Baptista
2-834-2-030	Gonçalo Joaquin Almeida Baptista
2-852-2-049	Gonçalo Joaquin Almeida Sá
2-845-2-059	Gonçalo Joaquin Almeida
2-844-2-041	Gonçalo Almeida Sá Baptista
2-836-2-132	Gonçalo Almeida Baptista

6.6 Illustrative Example Linkages

In order to clarify the preceding sections on linkage variables, criteria, and particularly procedure, two illustrative example linkages are drawn from the Viana data. The first example is taken from Stage 5A, where a standardised full name consisting of six standardised component names (António Pereira Cirne Sá/Silva Besera Fagundes) has been selected from a chain as the first standardised full name with the largest number of component names. The chain from which this block has been selected is one of four long chains comprising many combinations of standardised component names beginning with the common first component names António, João, José, and Manuel, these chains contain 345, 135, 356, and 253 blocks, 999, 274, 807, and 725 sub-blocks, and 2781, 920, 2516, and 2120 records, respectively.

In Table 12, the ordering of blocks by some measure of the likelihood that members of the block match members of the master block is shown. In Table 13, the division of blocks into sub-blocks is illustrated. Finally, in Table 14a, the extracted records of one historical person are shown, and in Table 14b, the records returned to the main chain are shown.

An important point that must be made is that, having extracted the records of one historical person, the remaining records of the sub-chain must be returned to the main chain, the linkage procedure must not continue to work on the remaining records of the sub-chain. The reason for this is best illustrated with a simple example. The remaining records in the sub-chain all have the standardised full name António Pereira Sá/Silva, this is a subset of António Joaquin-Pereira Sá/Silva, which is the standardised full name of a block still in the main chain.

An examination of the records presented in Tables 14a and 14b suggests that the record linkage is very satisfactory. In particular, it can be seen that if the

first record of block 6 sub-block 1 (NID 2-844-2-163) had been individually compared with the master block, it would have been linked to it. The record was not linked to the master block because a more likely link exists between it and the other record of block 6 sub-block 1 (NID 2-845-2-214), that link being more likely because the standardised full names are identical (also there is a parish match and a place of birth match)

The second illustrative example linkage is chosen to demonstrate the limitations of the record linkage procedure adopted. This example concerns the sub-blocking of just one block – the block with standardised full name Manuel José Rodrigues. This block was alluded to earlier in the description of Stage 4 of the linkage strategy. The linkage procedure divides the 99 records of this block among eight sub-blocks.

In Table 15, the resultant sub-blocks are presented, Table 15a contains the records of the first sub-block, and Table 15b contains the other sub-blocks. On the one hand, an examination of the records presented in Table 15a suggests that the linkage of the first sub-block is not very satisfactory. In particular, it can be seen that in many years in the 1850s three electors are included in the sub-block, so at least three different historical persons are contained in the first sub-block. On the other hand, an examination of the records presented in Table 15b suggests that the linkage for the remaining sub-blocks is quite satisfactory. The second sub-block contains the records of a priest, the third contains the records of a wood-carver, and, the fourth contains the records of a shoemaker. Finally, The last four sub-blocks each contain just one single cemetery list record, it is likely that one or two of these may belong to records which have been allocated to the first sub-block.

The problem occurs because the linkage procedure is unable to distinguish adequately between the records which are allocated to the first sub-block. Considering the information available from the electoral registers of 1851, the characteristics of the three individuals are as follows

Parish	EYB	MSC	OFC	OSC	Actual Occupation
Santa Maria Maior	1805	7	6	3	<i>Vendeiro</i> (Trader)
Monsserate	1804	3	4	3	<i>Pintor</i> (Painter)
Monsserate	1807	7	4	2	<i>Sapateiro</i> (Shoemaker)

As can be seen, two of the three electors live in the same parish, all three have estimated years of birth within three years of each other, two are married, two are engaged in trades, and two are engaged in the tertiary occupation sector. Of course, it would be possible to adjust the linkage procedure so that these records would be separated. But, there are many instances where linkage variables exhibit legitimate change, so, if the linkage procedure were adjusted, other sub-blocks would be incorrectly partitioned.

Table 12. Illustrative Example Linkage 1:**Stage 3 Ordering of Blocks (B)**

NID	B	Standardised Full Name
2-856-1-048	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-861-1-054	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-862-1-064	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-863-1-073	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-864-1-071	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-865-1-068	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-866-1-061	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-867-1-065	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-868-1-067	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-869-1-064	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-870-1-064	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-871-1-062	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-872-1-066	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-874-1-062	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-875-1-060	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-876-1-062	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-877-1-071	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-878-1-071	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-880-1-101	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-881-1-097	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-883-1-118	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-888-1-112	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-891-1-125	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-894-1-078	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-895-1-068	1	António Pereira Cirne Sá/Silva Besera Fagundes
4-900-0-136	1	António Pereira Cirne Sá/Silva Besera Fagundes
2-873-1-065	2	António Cirne Sá/Silva Besera Fagundes
2-845-1-035	3	António Pereira Sá/Silva Besera Fagundes
2-847-1-027	4	António Pereira Besera Fagundes
2-846-1-025	5	António Pereira Cirne
2-852-1-038	5	António Pereira Cirne
2-853-1-046	5	António Pereira Cirne
2-854-1-051	5	António Pereira Cirne
2-855-1-048	5	António Pereira Cirne
2-857-1-041	5	António Pereira Cirne
2-858-1-052	5	António Pereira Cirne
2-859-1-045	5	António Pereira Cirne
2-860-1-053	5	António Pereira Cirne
2-844-2-163	6	António Pereira Sá/Silva
2-845-2-214	6	António Pereira Sá/Silva
2-855-2-026	6	António Pereira Sá/Silva
3-867-0-010	6	António Pereira Sá/Silva
2-911-2-169	6	António Pereira Sá/Silva
4-919-0-101	6	António Pereira Sá/Silva

Table 13. Illustrative Example Linkage 1:

Stage 4 Division of Blocks (B) into Sub-Blocks (SB)									
Linkage Variables									
NID	B	SB	BPC	EYB	MSC	OFC	OSC	NAME	ACRONYM
2-856-1-048	1	1		1822	2	10	3	APCSBF	
2-861-1-054	1	1		1815	2	10	3	APCSBF	
2-862-1-064	1	1		1827	7	10	3	APCSBF	
2-863-1-073	1	1		1827	7			APCSBF	
2-864-1-071	1	1		1827	7	10	3	APCSBF	
2-865-1-068	1	1		1827	7	10	3	APCSBF	
2-866-1-061	1	1		1827	7	10	3	APCSBF	
2-867-1-065	1	1		1827	7	10	3	APCSBF	
2-868-1-067	1	1		1827	7	10	3	APCSBF	
2-869-1-064	1	1		1827	7	10	3	APCSBF	
2-870-1-064	1	1		1827	7	10	3	APCSBF	
2-871-1-062	1	1		1827	7	10	3	APCSBF	
2-872-1-066	1	1		1827	7	10	3	APCSBF	
2-874-1-062	1	1		1827	7	10	3	APCSBF	
2-875-1-060	1	1		1827	7	10	3	APCSBF	
2-876-1-062	1	1		1827	7	10	3	APCSBF	
2-877-1-071	1	1		1827	7	10	3	APCSBF	
2-878-1-071	1	1		1827	7	10	3	APCSBF	
2-880-1-101	1	1		1828	7	10	3	APCSBF	
2-881-1-097	1	1		1828	7	10	3	APCSBF	
2-883-1-118	1	1		1827	7	10	3	APCSBF	
2-888-1-112	1	1		1827	7	10	3	APCSBF	
2-891-1-125	1	1		1827	7	10	3	APCSBF	
2-894-1-078	1	1		1828	7	10	3	APCSBF	
2-895-1-068	1	1		1828	7	10	3	APCSBF	
4-900-0-136	1	1	11010101011	1819	2	10	3	APCSBF	
2-873-1-065	2	1		1827	7	10	3	APCSBF	
2-845-1-035	3	1		1819	6	10	3	APCSBF	
2-847-1-027	4	1		1821	1	10	3	APCSBF	
2-846-1-025	5	1		1820	1	10	3	APC	
2-852-1-038	5	1	11010101011	1818	7	10	3	APC	
2-853-1-046	5	1		1822	7	10	3	APC	
2-854-1-051	5	1		1821	2	10	3	APC	
2-855-1-048	5	1		1821	2	10	3	APC	
2-857-1-041	5	1		1823	2	10	3	APC	
2-858-1-052	5	1		1824		10	3	APC	
2-859-1-045	5	1		1824	2	10	3	APC	
2-860-1-053	5	1		1825	2	10	3	APC	
2-844-2-163	6	1	11000401000	1819		6	3	AP S	
2-845-2-214	6	1	11000401000	1820		7	3	AP S	
2-855-2-026	6	2		1800	7	6	3	AP S	
3-867-0-010	6	3	11010101010	1854			3	AP S	
2-911-2-169	6	4		1883	3	4	2	AP S	
4-919-0-101	6	4	11010102000	1874	2	4	2	AP S	

Table 14a. Illustrative Example Linkage 1:**Stage 5 Extracted Historical Persons****Linkage Variables**

NID	B	SB	BPC	EYB	MSC	OFC	OSC	NAME	ACRONYM
2-845-1-035	3	1		1819	6	10	3	AP SBF	
2-846-1-025	5	1		1820	1	10	3	APC	
2-847-1-027	4	1		1821	1	10	3	AP BF	
2-852-1-038	5	1	11010101011	1818	7	10	3	APC	
2-853-1-046	5	1		1822	7	10	3	APC	
2-854-1-051	5	1		1821	2	10	3	APC	
2-855-1-048	5	1		1821	2	10	3	APC	
2-856-1-048	1	1		1822	2	10	3	APCSBF	
2-857-1-041	5	1		1823	2	10	3	APC	
2-858-1-052	5	1		1824		10	3	APC	
2-859-1-045	5	1		1824	2	10	3	APC	
2-860-1-053	5	1		1825	2	10	3	APC	
2-861-1-054	1	1		1815	2	10	3	APCSBF	
2-862-1-064	1	1		1827	7	10	3	APCSBF	
2-863-1-073	1	1		1827	7			APCSBF	
2-864-1-071	1	1		1827	7	10	3	APCSBF	
2-865-1-068	1	1		1827	7	10	3	APCSBF	
2-866-1-061	1	1		1827	7	10	3	APCSBF	
2-867-1-065	1	1		1827	7	10	3	APCSBF	
2-868-1-067	1	1		1827	7	10	3	APCSBF	
2-869-1-064	1	1		1827	7	10	3	APCSBF	
2-870-1-064	1	1		1827	7	10	3	APCSBF	
2-871-1-062	1	1		1827	7	10	3	APCSBF	
2-872-1-066	1	1		1827	7	10	3	APCSBF	
2-873-1-065	2	1		1827	7	10	3	APCSBF	
2-874-1-062	1	1		1827	7	10	3	APCSBF	
2-875-1-060	1	1		1827	7	10	3	APCSBF	
2-876-1-062	1	1		1827	7	10	3	APCSBF	
2-877-1-071	1	1		1827	7	10	3	APCSBF	
2-878-1-071	1	1		1827	7	10	3	APCSBF	
2-880-1-101	1	1		1828	7	10	3	APCSBF	
2-881-1-097	1	1		1828	7	10	3	APCSBF	
2-883-1-118	1	1		1827	7	10	3	APCSBF	
2-888-1-112	1	1		1827	7	10	3	APCSBF	
2-891-1-125	1	1		1827	7	10	3	APCSBF	
2-894-1-078	1	1		1828	7	10	3	APCSBF	
2-895-1-068	1	1		1828	7	10	3	APCSBF	
4-900-0-136	1	1	11010101011	1819	2	10	3	APCSBF	

Table 14b. Illustrative Example Linkage 1:

Stage 5 Records Returned to Main Chain											
Linkage Variables											
NID	B	SB	BPC	EYB	MSC	OFC	OSC	NAME	ACRONYM		
2-844-2-163	6	1	11000401000	1819		6	3	A P	S		
2-845-2-214	6	1	11000401000	1820		7	3	A P	S		
2-855-2-026	6	2		1800	7	6	3	A P	S		
3-867-0-010	6	3	11010101010	1854				A P	S		
2-911-2-169	6	4		1883	3	4	2	A P	S		
4-919-0-101	6	4	11010102000	1874	2	4	2	A P	S		

æ

6.7 Performance of the Approaches Adopted

No absolute measure of the correctness of the record linkage of the Viana data has yet been developed. Such a measure would ideally need to consider the effects of the coding, standardisation, and linkage procedures separately. Nevertheless, while to guesstimate the value of such a measure would merely provide spurious accuracy, the micro-analysis presented in Sections 6.3 and 6.4 do suggest that a figure such as the proportion of records linked correctly might be extremely high.

It is felt that by initially proceeding without the aid of sophisticated record linkage programs, the problems inherent in the Viana data have been closely understood, in this respect, the initial linkage of the muster-rolls proved invaluable. It is also felt that, as a result, the software and techniques required to overcome these problems and link the data both automatically and accurately, have been more effectively developed.

Comparing the techniques presented here with those reviewed earlier, several points are worth emphasising. First, the use of a database system has proved invaluable, more particularly, the use of the Scientific Information Retrieval (SIR) Data Base Management System (DBMS) for the storage, coding, linkage, and subsequent retrieval of the Viana data has proved invaluable. Second, although the Viana data can certainly not be described as being of high quality, the standardisation and coding process allows the individual standardisation of variables to the extent required for record linkage, any and all standardisation and coding problems can be overcome. Third, another advantage of the record linkage procedure is that the usual assumption that the likelihood of any randomly chosen pair of records pertaining to the same historical person is independent of information available elsewhere is unnecessary. By using chains and sub-chains of blocks and sub-blocks, the record linkage of clusters of records becomes possible.

On the other hand, there are two apparent disadvantages in the application of the techniques presented here. First, as demonstrated with the second illustrative

Table 15a. Illustrative Example Linkage 2:**Stage 4 Division of Blocks (B) into Sub-Blocks (SB)****Linkage Variables**

CHAIN	B	SB	NID	BPC	EYB	MSC	OFC	OSC
10100925	152	1	1-810-0-804		1805	2	4	2
10100925	152	1	2-837-2-105					
10100925	152	1	2-839-2-101				7	3
10100925	152	1	2-844-1-290	11010103010	1804	2	4	3
10100925	152	1	2-845-1-364		1807	2	4	3
10100925	152	1	2-848-1-250		1806	3	4	3
10100925	152	1	2-849-1-284		1814	2	4	2
10100925	152	1	2-849-1-324		1807	3	4	3
10100925	152	1	2-851-1-378		1805	7	6	3
10100925	152	1	2-851-2-162		1804	3	4	3
10100925	152	1	2-851-2-184		1807	7	4	2
10100925	152	1	2-852-2-111	11010103010	1807	3	4	3
10100925	152	1	2-853-1-332		1806	7	6	3
10100925	152	1	2-853-2-135		1813	3	4	3
10100925	152	1	2-853-2-158		1809	7	4	2
10100925	152	1	2-854-1-309		1805	7	6	3
10100925	152	1	2-854-2-144		1806	3	5	3
10100925	152	1	2-854-2-145		1808	2	4	2
10100925	152	1	2-855-1-299		1805	7	6	3
10100925	152	1	2-855-2-140		1806	3	4	3
10100925	152	1	2-855-2-141		1808	2	4	2
10100925	152	1	2-856-1-296		1806	7	6	3
10100925	152	1	2-856-2-138		1807	3	4	3
10100925	152	1	2-857-1-273		1807	7	6	3
10100925	152	1	2-857-2-132		1808	3	4	3
10100925	152	1	2-858-1-312		1808	2	6	3
10100925	152	1	2-858-2-140		1809	3	4	3
10100925	152	1	2-859-1-291		1808	7	6	3
10100925	152	1	2-859-2-141		1810	1	4	3
10100925	152	1	2-860-1-326		1808	2	6	3
10100925	152	1	2-861-1-322		1799	2	6	3
10100925	152	1	2-862-1-345		1799	2	6	3
10100925	152	1	2-863-1-439		1799	2	6	3
10100925	152	1	2-864-1-406		1799	2	6	3
10100925	152	1	2-865-1-407		1799	2	6	3
10100925	152	1	2-866-1-397		1799	2	6	3
10100925	152	1	2-867-1-420		1799	2	6	3
10100925	152	1	2-868-1-440		1799	2	6	3
10100925	152	1	2-869-1-427		1799	2	6	3
10100925	152	1	2-870-1-427		1799	2	6	3
10100925	152	1	2-871-1-418		1799	2	6	3
10100925	152	1	2-872-1-456		1799	2	6	3
10100925	152	1	2-873-1-437		1800	2	6	3
10100925	152	1	2-874-1-436		1800	2	6	3
10100925	152	1	2-875-1-431		1800	2	6	3
10100925	152	1	2-876-1-432		1800	7	6	3
10100925	152	1	2-877-1-441		1800	7	6	3
10100925	152	1	2-878-1-455		1800	7	6	3
10100925	152	1	2-880-1-645		1800	2	6	3
10100925	152	1	2-881-1-624		1800	2	6	3
10100925	152	1	4-882-0-028	11010102220	1808	2	6	3

Table 15b. Illustrative Example Linkage 2:

Stage 4 Division of Blocks (b) into Sub-Blocks (SB)									
Linkage Variables									
CHAIN	B	SB	NID	BPC	EYB	MSC	OFC	OSC	
10100925	152	2	2-834-1-187						
10100925	152	2	2-835-1-161		1792	4	9	3	
10100925	152	2	2-837-1-226		1793	4	9	3	
10100925	152	2	2-838-1-248						
10100925	152	2	2-839-1-225						
10100925	152	2	2-845-1-331						
10100925	152	2	2-848-1-239		1799	4	9	3	
10100925	152	2	2-849-1-314		1796		9	3	
10100925	152	2	2-853-1-358		1797	3	9	3	
10100925	152	2	2-854-1-310		1803	4	9	3	
10100925	152	2	2-855-1-300		1816	4	9	3	
10100925	152	2	2-856-1-297		1816	4	9	3	
10100925	152	2	2-857-1-274		1807	4	9	3	
10100925	152	2	2-858-1-313		1808	4	9	3	
10100925	152	2	2-859-1-292		1809	4	9	3	
10100925	152	2	2-860-1-325		1809	4	9	3	
10100925	152	2	2-861-1-323		1808	4	9	3	
10100925	152	2	2-862-1-344		1808	4	9	3	
10100925	152	2	2-863-1-438		1808	4	9	3	
10100925	152	2	2-864-1-405		1808	4	9	3	
10100925	152	2	2-865-1-406		1808	4	9	3	
10100925	152	2	2-866-1-396		1808	4	9	3	
10100925	152	2	2-867-1-419		1808	4	9	3	
10100925	152	2	2-868-1-439		1808	4	9	3	
10100925	152	2	2-869-1-426		1808	4	9	3	
10100925	152	2	2-870-1-426		1808	4	9	3	
10100925	152	2	2-871-1-417		1808	4	9	3	
10100925	152	2	2-872-1-455		1808	4	9	3	
10100925	152	2	2-873-1-436		1808	4	9	3	
10100925	152	2	2-874-1-435		1808	4	9	3	
10100925	152	2	2-875-1-430		1808	4	9	3	
10100925	152	2	4-875-0-044	11010101010	1808	4	9	3	
10100925	152	3	2-877-1-534				9	3	
10100925	152	3	2-880-1-646		1837	2	8	3	
10100925	152	3	2-881-1-625		1846	1	4	2	
10100925	152	3	2-883-1-710		1846	1	4	2	
10100925	152	3	2-888-1-630		1846	6	4	2	
10100925	152	3	2-891-1-690		1846	2	4	2	
10100925	152	3	2-895-1-335		1845	2	4	2	
10100925	152	3	4-906-0-089	11010101011	1839	6	4	2	
10100925	152	4	2-880-2-439		1842	3	9	2	
10100925	152	4	2-881-2-446		1854	7	4	2	
10100925	152	4	2-883-2-453		1854	7	4	2	
10100925	152	4	3-884-0-012	11010101011	1854	7	4	2	
10100925	152	5	4-866-0-059	11010102300	1857	2	4	2	
10100925	152	6	4-870-0-014			1	7	3	
10100925	152	7	4-878-0-058	11010101011		2	8	3	
10100925	152	8	4-883-0-119	11010101012	1858	2			

example linkage of Section 6.6, the linkage procedure is designed so that there is sometimes a slight tendency toward over-linkage. This means that where several records share very similar characteristics (e.g. those of the three electors with standardised full name Manuel José Rodrigues), they are simply linked together. It is argued that this design feature of the linkage procedure is not a great disadvantage. In general, slight over-linkage is preferable to under-linkage, it is reassuring to know that all the records of a particular historical person are linked together. Also, the effects of slight over-linkage can be controlled in analyses of the linked data. Where a slightly over-linked individual life history is analysed, the characteristics of the constituent records can be chosen to be an unbiased sample of the available characteristics.

The second disadvantage of the linkage procedure is that it may appear to be somewhat ad hoc – although the match scoring functions have been tuned to produce the desired linkage of the Viana data, they are not directly applicable elsewhere. Nevertheless, the methodology and general approach adopted are very flexible, and would therefore only require some modifications of detail in order that they might be applied elsewhere.

Finally, it is again noted that the record linkage process presented here may require some further development before a full reconstitution of the whole population of Viana, incorporating the vital registration records can be attempted.

7. SUMMARY

This paper has addressed a number of methodological and substantive issues in record linkage in general, and in the reconstitution of the Town/City of Viana in particular.

First, the economic and demographic history of Viana were discussed, and local and national events in the nineteenth century were summarised in order to establish the circumstances under which the sources being used were created. Then, the manuscript sources (muster-rolls, electoral registers, passport books, and cemetery lists) on which the reconstitution is currently based were described in detail.

Reconstitution methodology was introduced with a review of some record linkage studies and techniques. Record linkage studies were classified according to whether or not variation or errors or both exist in identifying items of information, and whether or not there is duplication of identifying item sets. The combination of these problems was seen to determine the approach to record linkage. Nevertheless, it was argued that whatever approach is adopted ought to be fully automatic, ensuring both that linkage criteria are carefully defined beforehand, and that those criteria are consistently applied.

Next, the use of the Scientific Information Retrieval (SIR) Data Base Man-

agement System (DBMS) for the storage, standardisation and coding, linkage, and subsequent retrieval of the Viana data was described

Finally, the record linkage of the Viana data was described in detail. The record linkage was not presented within a formal framework because to do so lies beyond the scope of this research, however, it is fair to say that the groundwork for such a framework has been laid

APPENDIX A.1

Example Muster-Roll

RUA 8/9							
Vila	Número	Nome	Estado	Idade	Profissão	Relações	Outros
		1 08 007 01					
7	A	1					
B		1					
		2					
C	1	1					
		1					
		1					
		4					
		2					
		3					

Example Electoral Register

Cidade de São Paulo de Santa Maria Maior					
Nomes	Idade	Estado	Profissão	Pendimentos anual (legua)	
João da Silva G. Abigail	118	25	Gratuito	100,0000	
João da Silva M. J.	120	40	Gratuito	100,0000	
João da Silva M. J.	120	50	Gratuito	200,0000	
João da Silva M. J.	120	60	Gratuito	100,0000	
João da Silva M. J.	120	50	Gratuito	100,0000	
João da Silva M. J.	120	40	Gratuito	100,0000	
João da Silva M. J.	120	50	Gratuito	100,0000	
João da Silva M. J.	120	50	Gratuito	100,0000	
João da Silva M. J.	120	40	Gratuito	200,0000	
João da Silva M. J.	120	60	Gratuito	100,0000	
João da Silva M. J.	120	36	Gratuito	100,0000	
João da Silva M. J.	120	49	Gratuito	100,0000	
João da Silva M. J.	120	43	Gratuito	100,0000	
João da Silva M. J.	120	55	Gratuito	100,0000	
João da Silva M. J.	120	70	Gratuito	100,0000	
João da Silva M. J.	120	60	Gratuito	200,0000	
João da Silva M. J.	120	50	Gratuito	100,0000	
João da Silva M. J.	120	43	Gratuito	100,0000	
João da Silva M. J.	120	46	Gratuito	100,0000	
João da Silva M. J.	120	39	Gratuito	100,0000	
João da Silva M. J.	120	36	Gratuito	100,0000	
João da Silva M. J.	120	36	Gratuito	100,0000	
João da Silva M. J.	120	37	Gratuito	100,0000	
João da Silva M. J.	120	62	Gratuito	100,0000	
João da Silva M. J.	120	38	Gratuito	100,0000	
João da Silva M. J.	120	60	Gratuito	200,0000	
João da Silva M. J.	120	53	Gratuito	200,0000	
João da Silva M. J.	120	36	Gratuito	100,0000	
João da Silva M. J.	120	61	Gratuito	100,0000	

APPENDIX A.3

Example Passport Book

3 856 0005	
GOVERNO CIVIL DE VIANNA DO CASTELLO	O GOVERNO CIVIL DO DISTRITO DE VIANNA DO CASTELLO
1ª REPARTIÇÃO	CONCEDE passaporte a <i>António da Silva</i>
Passaporte do exterior N.º 838	filho de <i>António da Silva</i> natural de <i>Viana do Castelo</i>
Signaes	residente em <i>Viana do Castelo</i>
Idade <i>14</i> annos	de profissão --
Altura <i>1.70</i>	para <i>obter a licença com a obrigação de</i>
Rosto <i>Redondo</i>	<i>para o serviço de Portugal, e para o serviço</i>
Cabello <i>Castanho</i>	<i>de Portugal, e para o serviço de Portugal, e para o serviço</i>
So brolho <i>Castanho</i>	<i>de Portugal, e para o serviço de Portugal, e para o serviço</i>
Olhos <i>Castanhos</i>	abonado por <i>esse</i>
Nariz <i>Regular</i>	<i>de Portugal, e para o serviço de Portugal, e para o serviço</i>
Boca <i>Regular</i>	E cumprido com a obrigação &c.
Côr <i>Castanho</i>	Valeoso por tempo de <i>60</i> dias para sair destes Reinos
Signaes particulares	Dado em Vianna aos <i>7</i> de <i>abril</i> de 185 <i>6</i>
<i>etc.</i>	O Governador Civil -
	Portador <i>António da Silva</i>
	Assignatura do Fielor <i>João da Silva</i>
	Duo <i>Gaspar João</i>

APPENDIX A.4

Example Cemetery List

Date		Time		Place		Remarks	
Year	Month	Day	Hour	Location	Altitude	Direction	Notes
1919	11	1	11:00	St. Peter's Church	4878	0050	
1919	11	2	11:00	St. Peter's Church	4878	0050	
1919	11	3	11:00	St. Peter's Church	4878	0050	
1919	11	4	11:00	St. Peter's Church	4878	0050	
1919	11	5	11:00	St. Peter's Church	4878	0050	
1919	11	6	11:00	St. Peter's Church	4878	0050	
1919	11	7	11:00	St. Peter's Church	4878	0050	
1919	11	8	11:00	St. Peter's Church	4878	0050	
1919	11	9	11:00	St. Peter's Church	4878	0050	
1919	11	10	11:00	St. Peter's Church	4878	0050	
1919	11	11	11:00	St. Peter's Church	4878	0050	
1919	11	12	11:00	St. Peter's Church	4878	0050	
1919	11	13	11:00	St. Peter's Church	4878	0050	
1919	11	14	11:00	St. Peter's Church	4878	0050	
1919	11	15	11:00	St. Peter's Church	4878	0050	
1919	11	16	11:00	St. Peter's Church	4878	0050	
1919	11	17	11:00	St. Peter's Church	4878	0050	
1919	11	18	11:00	St. Peter's Church	4878	0050	
1919	11	19	11:00	St. Peter's Church	4878	0050	
1919	11	20	11:00	St. Peter's Church	4878	0050	
1919	11	21	11:00	St. Peter's Church	4878	0050	
1919	11	22	11:00	St. Peter's Church	4878	0050	
1919	11	23	11:00	St. Peter's Church	4878	0050	
1919	11	24	11:00	St. Peter's Church	4878	0050	
1919	11	25	11:00	St. Peter's Church	4878	0050	
1919	11	26	11:00	St. Peter's Church	4878	0050	
1919	11	27	11:00	St. Peter's Church	4878	0050	
1919	11	28	11:00	St. Peter's Church	4878	0050	
1919	11	29	11:00	St. Peter's Church	4878	0050	
1919	11	30	11:00	St. Peter's Church	4878	0050	
1919	11	31	11:00	St. Peter's Church	4878	0050	

APPENDIX B*Data Recorded on the Electoral Registers**Key*

Variable	Description
1	Name
2	Age in Years
3	Marital Status
4	Occupation
5	Birthplace
6	Road of Residence
7	Literacy
8	Household Head
9	Number of Tax Categories Recorded
10	Social Qualifications
11	Observations
12	Eligible for Election
13	Eligible for Jury Service
14	Major Contributor (one of the 40 highest taxpayers)

The final column (15) provides an indication of the quality of the document (A, B, or C) and whether it appears to be an update of the previous electoral register (U)

Year	Variable														
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1834	*	*	*	*		*	*	*	1	*	*				A
1835	*	*	*	*		*	*	*	1	*	*				A
1836	*	*	*	*		*	*	*	1	*	*				A
1837	*														A
1838	*														A
1839	*														A
1840	*	*		*			*		4						A
1841	*	*		*			*		4						B
1842	*	*	*	*	*		*		4						A
1843	*	*	*	*	*		*		7						B
1844	*	*	*	*	*		*		5						C
1845	*	*	*	*			*		4						C
1846	*	*	*	*			*		4						C
1847	*	*	*	*			*		4						C
1848	*	*	*	*			*		4						B
1849	*	*	*	*			*		4						B
1850	*	*	*	*			*		4						B
1851	*	*	*	*			*		4						A
1852	*	*	*	*	*		*		4						C
1853	*	*	*	*		*	*		6		*				B
1854	*	*	*	*		*			4	*	*	*			B
1855	*	*	*	*		*			4	*	*	*			A
1856	*	*	*	*		*			5	*	*	*	*		B
1857	*	*	*	*		*			5	*	*	*	*	*	B
1858	*	*	*	*		*			5	*	*	*	*	*	B
1859	*	*	*	*		*			5	*	*	*	*	*	A
1860	*	*	*	*		*			5	*	*	*	*	*	B
1861	*	*	*	*		*			5	*	*	*	*	*	A
1862	*	*	*	*		*			5	*	*	*	*	*	A
1863	*	*	*	*		*			5	*	*	*	*	*	A
1864	*	*	*	*		*			5	*	*	*	*	*	A
1865	*	*	*	*		*			5	*	*	*	*	*	A
1866	*	*	*	*		*			5	*	*	*	*	*	A
1867	*	*	*	*		*			5	*	*	*	*	*	A
1868	*	*	*	*		*			5	*	*	*	*	*	A
1869	*	*	*	*		*			5	*	*	*	*	*	B
1870	*	*	*	*		*			5	*	*	*	*	*	A
1871	*	*	*	*		*			5	*	*	*	*	*	A
1872	*	*	*	*		*			5	*	*	*	*	*	AU
1873	*	*	*	*		*			5	*	*	*	*	*	B
1874	*	*	*	*		*			5	*	*	*	*	*	BU
1875	*	*	*	*		*			5	*	*	*	*	*	BU
1876	*	*	*	*		*			5	*	*	*	*	*	AU
1877	*	*	*	*		*			5	*	*	*	*	*	BU
1878	*	*	*	*		*			5	*	*	*	*	*	BU
1880	*	*	*	*		*			5	*	*	*	*	*	A
1881	*	*	*	*		*			2	*					B
1883	*	*	*	*		*			2	*					AU
1888	*	*	*	*		*			2	*					B
1891	*	*	*	*		*				*					A
1894	*	*	*	*		*				*					A
1895	*	*	*	*		*			1						A
1911	*	*	*	*	*										A
1931	*	*	*	*											A

(an * denotes that the information was recorded)

MANUSCRIPT SOURCES*Arquivo Distrital de Viana do Castelo**Baptismos**Santa Maria Maior, -1876**Monsserrate, -1867**Casamentos**Santa Maria Maior, -1872**Monsserrate, -1867**Arquivo do Governador Civil do Distrito de Viana do Castelo**Livros dos Passaportes, 1835-1935**Estatísticas da População do Distrito de Vianna do Castello, 1837-1867 and 1869-1890 (excluding 1863 and 1877, years preceding those in which national censuses were conducted)**Registo da Lista Geral dos individuos do Concelho de Vianna para votar na Eleição da Câmara Municipal, 1888**Arquivo Municipal de Viana do Castelo**Lista Geral das Companhias de Ordenanças do Distrito de Vianna da Terceira Brigada**Registo da Lista Geral dos individuos do Concelho de Vianna para votar na Eleição da Câmara Municipal 1834-1840, 1842-1847, 1849-1850, 1852-1853, 1855, 1869, 1881, 1891, 1894, 1931**Registo dos menores sepultados no cemitério publico de Vianna do Castello, 1855-1930**Registo dos adultos sepultados no cemitério publico de Vianna do Castello, 1855-1930**Arrolamento das Pessoas e Coisas, 1871**Biblioteca Municipal de Viana do Castelo**Registo da Lista Geral dos individuos do Concelho de Vianna para votar na Eleição da Câmara Municipal 1841, 1848, 1851, 1854, 1856-1868, 1870-1878, 1880, 1883*

PRINTED SOURCES

- Abbreviatura de Endereços* (Telephone Directory), 1986 Viana do Castelo
Almanak do Distrito de Vianna do Castello Biblioteca Municipal de Viana do Castelo
A Aurora do Lima, 1862, 1880, 1882-1884, 1886-1888, 1890, 1895 Biblioteca Municipal de Vianna do Castelo
A Aurora do Lima, 1856-1870, 1895-1951 Biblioteca Municipal do Porto
Censo da População do Reino de Portugal 1/1/1864, 1/1/1878, 1/12/1890, 1/12/1900 Lisboa
Denominação de Ruas, 1937 Arquivo Municipal de Viana do Castelo, No 1226
Grande Enciclopédia Portuguesa e Brasileira Lisboa e Rio de Janeiro Editorial Enciclopédia
Legislação Portuguesa, 1840, 1842, 1852, 1859, 1883 British Museum
O Primeiro de Janeiro, 1882 Lisboa

REFERENCES

- Acheson, E D (Ed) (1968) *Record Linkage in Medicine* Edinburgh and London Livingstone
 Åkerman, Sune (1977) An Evaluation of the Family Reconstitution Technique *Scandinavian Economic History Review*, 25 160-170
 Åkerman, Sune (1981) How Did the Great Decline in Fertility Start? A Study Based on Retrospective Interviews In Brandstrom, A & Sundin, J (1981) *Tradition and Transition Studies in Microdemography and Social Change* Report No 2 from the Demographic Data Base, University of Umeå, Sweden
 Alarcão, Alberto de (1983) *The Portuguese Population in the History of Civilizations An introductory analysis for its comprehension, and some problems* Centro de Estudos de Economia Agrária, Instituto Gulbenkian de Ciência, Oeiras
 Armstrong, W A (1972) The Use of Information about Occupations In Wrigley, E Anthony, (Ed) (1972) *Nineteenth Century Society* Cambridge University Press

- Blayo, Ives (1973) Name Variations in a Village in Brie, 1750–1860 In Wringley, E Anthony (Ed) (1973) *Identifying People in the Past* London Edward Arnold
- Bouchard, Gérard (1986) The Processing of Ambiguous Links in Computerised Family Reconstruction *Historical Methods*, 19 (1) 9–19
- Braga, Custódio Capela (1985) Para a História de Viana do Castelo visitas de sade aos barcos acostados no Cabedelo da Foz do Lima de 1720 a 1774 *Cadernos Vianenses*, 9 95–117
- Brandão, Maria de Fátima & Rowland, Robert (1980) História da Propriedade e Comunidade Rural Questões de Método *Análise Social*, Vol 16, Nos 61–62 173–207
- Brandão, Maria de Fátima & Rui Graça de Castro Feijó (1984). Entre Textos e Contextos os Estudos de Comunidade e as suas Fontes Históricas *Análise Social*, Vol 20, No 83 489–503
- Brandstrom, A & Sundin, J (1981) *Tradition and Transition Studies in Microdemography and Social Change* Report No 2 from the Demographic Data Base, University of Umeå, Sweden
- Caldas, José (1919) *História de um Fogo-morto (Subsídios para uma História Nacional) 1258–1848* Porto Renascença Portuguesa.
- Castro, Francisco Cyrne de (1954) A Cólera-morbus no Distrito de Viana do Castelo *Arquivo do Alto Minho*, 4 106–119
- Castro, Francisco Cyrne de (1955) Ainda “A Cólera-morbus no Distrito de Viana do Castelo” *Arquivo do Alto Minho*, 4 106–119
- Castro, Francisco Cyrne de (1979) Viana no Comércio do Mar Algumas Notas *Cadernos Vianenses*, 2 18–21
- Coelho, Eusébio Candido C P Furtado (1861) *Estatística do Distrito de Vianna do Castello* Lisboa Imprensa Nacional
- Crespo, José (1957) *Monografia de Viana do Castelo* Viana do Castelo Câmara Municipal
- Danell, Christina (1981) The Demographic Data Base at Umeå University In Brandstrom, A & Sundin, J (1981) *Tradition and Transition Studies in Microdemography and Social Change* Report No 2 from the Demographic Data Base, University of Umeå, Sweden
- Denley, Peter & Deian Hopkin (Eds) (1987) *History and Computing* Manchester University Press
- Denley, Peter, Stefan Fogelvik & Charles Harvey (Eds) (1989) *History and Computing*, volume 2. Manchester University Press
- Dijk, Henk van (1977) Longitudinal Cohort Analysis of the Population of Eindhoven [Netherlands] In Sundin, Jan & Erik Soderland (Eds) (1979) *Time, Space and Man Essays on Microdemography* Report from the symposium Time, Space and Man in Umeå, 8–11 June 1977 Uppsala
- Doulton, David & Arno Kitts (1989) The Storing and Processing of Historical Data In Denley, Peter, Stefan Fogelvik & Charles Harvey (Eds) (1989) *History and Computing*, volume 2 Manchester University Press
- Drake, M. (1971) The Mid-Victorian Voter *Journal of Interdisciplinary History*, 1

- Du Bois Junior, N S d'Andrea (1965). A Document Linking Program for Digital Computers *Behavioural Science*, 10: 312-319
- Du Bois Junior, N S d'Andrea (1969) A Solution to the Problem of Linking Multivariate Records *Journal of the American Statistical Association*, 64: 163-174.
- Feijó, Rui Graça de Castro (1983) *Liberal Revolution, Social Change and Economic Development the region of Viana (NW Portugal) in the first three quarters of the nineteenth century* Unpublished Ph D thesis Oxford University
- Feijó, Rui Graça de Castro (1987) Um Exercício sobre Nomes *Boletín de la Asociacin de Demografia Histórica*, Año 5, No 1 50-63
- Felgueiras, Antero A M. (1979) A Ponte de Madeira *Cadernos Vianenses*, 2 45-49
- Felligi, Ivan P & Alan B Sunter (1967) An Optimal Theory for Record Linkage *Proceedings of the International Symposium on Automation of Population Registration Systems*, Vol 1 Jerusalem
- Felligi, Ivan P & Alan B Sunter (1969) A Theory for Record Linkage *Journal of the American Statistical Association*, 64 1138-1210
- Gauvreau, Danielle (1986) *The Study of Migration A Method Based on Re-constituted Families from Quebec Registers* Paper presented at the annual meeting of the Population Association of America, April, San Francisco
- Godinho, José de Magalhães (1969) *A Legislação Eleitoral e a sua Crítica* Lisboa Prelo
- Guerra, Luís de Figueiredo da (1880) *História da Cidade de Vianna* Arquivo Municipal de Viana do Castelo, Ms 709
- Halpern Pereira, Miriam (1971) *Livre Câmbio e Desenvolvimento Económico Portugal na Segunda Metade do Século XIX* Lisboa Edições Cosmos
- Herculano de Carvalho, Alexandre (1978) *Opúsculos*, I Lisboa
- Herlihy, David (1973) Problems of Record Linkages in Tuscan Fiscal Records of the fifteenth century In Wrigley, E Anthony (Ed) (1973) *Identifying People in the Past* London. Edward Arnold
- Hershberg, Theodore, Alan Burstein & Robert Dockhorn (1976) Record Linkage *Historical Methods Newsletter*, 9 137-163
- Hodson, F R, D G Kendall & P Tăutu (1971) (Eds) *Mathematics in the Archeological and Historical Sciences Proceedings of the Anglo-Romanian Conference*, Mamaia, 1970 Edinburgh University Press
- Howes, H W (1950) *The Gibraltarian The Origin and Development of the Population of Gibraltar from 1704*.
- Hubbard, M R & J E Fisher (1968) A Computer System for Medical Record Linkage In Acheson, E D (Ed) (1968) *Record Linkage in Medicine* Edinburgh and London Livingstone
- Iverson, K E (1962) *A Programming Language* New York Wiley
- Katz, M B (Ed) (1970) *The Hamilton Project*, Vol 2 Toronto Ontario Institute for Studies in Education
- Katz, M B & Tiller, John (1972) Record Linkage for Everyman A Semi-Automated Process *Historical Methods Newsletter*, 7

- Kelley, R, M H Skolnick & N Yasuda (1972) A Combinatorial Problem in Linking Historical Records *Historical Methods Newsletter*, 6 10-16
- Kitts, Arno, David Doulton, Ian Diamond & Elizabeth Reis (1987) Using the Database Management System SIR to Link Political Data from Viana do Castelo, Minho, Portugal, 1827-95 In Denley, Peter & Deian Hopkin (Eds) (1987) *History and Computing* Manchester University Press
- Kitts, Arno (1988) *An Analysis of the Components of Migration Viana do Castelo, Minho, 1826-1931* Unpublished Ph D thesis University of Southampton
- Légaré, J (1972) The Early Canadian Population Problems in Automatic Record Linkage *The Canadian Historical Review*, 53 424-439
- Lencastre, Francisco de (1869). *Índice Remissivo da Legislação Novíssima de Portugal, compreendendo os annos de 1833 até 1868* Lisboa Typographia Universal
- Levine, David (1976) The Reliability of Parochial Registration and the Representativeness of Family Reconstitution *Population Studies*, 30 (1) 107-122
- Livermore, H V (1966) *A New History of Portugal* Cambridge University Press
- Loureiro, Adolfo (1923) *Portos Marítimos de Portugal* Lisboa
- Macfarlane, Alan, S Harrison & C Jardine (1977) *Reconstructing Historical Communités* Cambridge University Press
- Mendes, José M. Amado (1980) Sobre as Relações Entre a Indústria Portuguesa e a Estrangeira no Século XIX *Análise Social*, Vol 16, Nos 61-62 31-52
- Moreira, Manuel António Fernandes (1984) *O Porto de Viana do Castelo na Época dos Descobrimentos* Viana do Castelo Câmara Municipal
- Nathan, G (1964) *On Optimal Matching Processes* Unpublished Ph.D thesis Case Institute of Technology, Cleveland, Ohio
- Nathan, G (1967) Outcome Probabilities for a Record Matching Process with Complete Invariant Information *Journal of the American Statistical Association*, 62 454-469
- Newcombe, Howard B, James M. Kennedy, S L Axford & A P James (1959) Automatic Linkage of Vital Records *Science*, 130 954
- Newcombe, Howard B & James M Kennedy (1962) Record Linkage Making Maximum Use of the Discriminating Power of Identifying Information *Communications of the Association for Computing Machinery*, 5 563-567
- Newcombe, Howard B. (1967) Record Linkage The Design of Efficient Systems for Linking Records into Individual and Family Histories *American Journal of Human Genetics*, 19 335-359
- Ní Bhrolcháin, M & I Timaeus (1983) *A General Approach to the Machine Handling of Event History Data with Special Reference to Employment Histories* CPS Research Paper No 83-1 London School of Hygiene & Tropical Medicine

- Nitzberg, D H (1968) Results of Research into the Methodology of Record Linkage In Acheson, E D (Ed) (1968) *Record Linkage in Medicine* Edinburgh and London Livingstone
- Nowell, Charles E (1952) *A History of Portugal* New York Van Nostrand
- Oliveira Martins, Joaquim Pedro de (1891) *Portugal Contemporâneo* Lisboa Guimarães Editores, 1953
- Peres, Damião (1935) O Setembrismo In *História de Portugal*, Vol 7 253–261 Barcelos Portucalense
- Peres, Damião (1935) *História de Portugal* Barcelos Portucalense
- Phillips Junior, William, Anita K Bahn & Mabel Miyasaki (1962) Person-Matching by Electronic Methods *Communications of the Association for Computing Machinery*, 5 404–407
- Pierson, Donald (1970) A Review of Sayers, Raymond S (Ed) (1968) *Portugal and Brazil in Transition* Minneapolis University of Minnesota Press *Journal of Inter-American Studies and World Affairs*, 12 455–465
- Pina-Cabral, João de (1986) *Sons of Adam, Daughters of Eve* Oxford Clarendon Press
- Pouyez, Christian, Raymond Roy & François Martin (1983) The Linkage of Census Name Data Problems and Procedures *Journal of Interdisciplinary History*, 14 (1). 129–152.
- Reis, Elizabeth de Azevedo (1987) *The Spatial Demography of Portugal in the Late Nineteenth Century Evidences from the 1864 and 1878 Censuses* Unpublished Ph D thesis University of Southampton
- Reis, Jaime (1984) O Atraso Económico Português em Perspectiva Histórica (1860–1913) *Análise Social*, Vol 20, No 80 7–28
- Roby, João F M Pinto (1846) *Exposição Analytica do Pronunciamento do Dia 17 de Maio em Braga e dos Actos da Junta Provisória nas Dias 17 e 18 do Dito Mês* Porto
- Rowland, Robert (1981) Âncora e Montaria, 1827 Duas Freguesias do Noroeste Segundo os Livros de Registo das Companhias de Ordenanças In *Estudos Contemporâneos*, 2/3, *Perspectivas sobre o Norte de Portugal* Porto
- Rowland, Robert (1987) [Verbal] Comments made on the discussion of "Methodological Issues in Record Linkage" presented by Arno Kitts at the Second Annual Conference of the Association for History and Computing, 20–22 March, Westfield College, University of London
- Ruiz, Joaquín del Moral (1980) A Independência Brasileira e a sua Repercussão no Portugal da Época (1810–34) *Análise Social*, Vol 16, No 64 779–795
- Runblom, Harald & Hans Norman (Eds) (1976) *From Sweden to America A History of the Migrations* Minneapolis University of Minnesota Press
- Russell, Bertrand (1912) *The Problems of Philosophy*, 11th impression Oxford University Press, 1983
- Sá, Vitor de (1969) *A Crise do Liberalismo e as Primeiras Manifestações das Ideias Socialistas em Portugal (1820–1852)* Lisboa

- Sampaio, Francisco (1981) *História Económica Análise Crítica à Tabela de Preços dos Cinco Mercados Reguladores (Viana, Caminha, Arcos, Monção, Ponte de Lima), Entre 1858-1868* Caminha
- Saraiva, José Hermano (1978) *História Concisa de Portugal* Lisboa
- Sayers, Raymond S (Ed) (1968) *Portugal and Brazil in Transition* Minneapolis University of Minnesota Press
- Schofield, Roger S (1972) Representativeness and Family Reconstitution *Annales de Démographie Historique* 121-125, Paris
- Serrão, Joel (1970) *Cronologia Geral da História de Portugal* Lisboa Iniciativas Editoriais
- Serrão, Joel (1973) *Fontes de Demografia Portuguesa, 1800-1862* Lisboa Livros Horizonte
- Serrão, Joel (1976) *Dicionário de História de Portugal* Lisboa Iniciativas Editoriais
- Silbert, A (1977) *Do Portugal de Antigo Regime ao Portugal Outocentista* Lisboa Livros Horizonte
- Silva, Domingos Oliveira (1985) *The Apogee and Decline of British Hegemony in Portugal 1807-1820* Unpublished Ph D thesis University of Southampton
- Skolnick, Mark H, A Moroni, C Cannings & L L Cavalli-Sforza (1970) The Reconstruction of Genealogies from Parish Books In Hodson, F R, D G Kendall & P Tăutu (1971) (Eds) *Mathematics in the Archaeological and Historical Sciences, Proceedings of the Anglo-Romanian Conference, Mamaia, 1970* Edinburgh University Press
- Skolnick, Mark H (1973) The Resolution of Ambiguities in Record Linkage In Wrigley, E Anthony (Ed) (1973) *Identifying People in the Past* London Edward Arnold
- Smith, A (1968). Preservation of Confidence at the Central Level In Acheson, E D (Ed) (1968) *Record Linkage in Medicine* Edinburgh and London Livingstone
- Smith, Adam (1776) *An Inquiry into the Nature and Causes of the Wealth of Nations* London Methuen, 1961
- Smith, Tom C, R Y Eng, & R T Lundy (1977) *Nakahara Family Planning and Population in a Japanese Village, 1717-1830* Stanford University Press
- Smythe, M (1968) Record Numbering In Acheson, E D (Ed) (1968) *Record Linkage in Medicine* Edinburgh and London Livingstone
- Sundin, Jan & Erik Soderland (Eds) (1979) *Time, Space and Man Essays on Microdemography*, Report from the symposium Time, Space and Man in Umeå, 8-11 June 1977 Uppsala
- Sundin, Jan (1984) *The Demographic Data Base at Umeå University Sources, Methods, and Products* Paper presented at the conference on Population Databases for Historical Research, August 13-17, Umeå, Sweden
- Sundin, Jan (1985) Preface of Winchester, Ian (1985) *Record Linkage in the Microcomputer Era A Survey*, Newsletter No 3 from the Demographic Data Base, University of Umeå, Sweden

- Sunter, A B (1968) A Statistical Approach to Record Linkage In Acheson, E D (Ed) (1968) *Record Linkage in Medicine* Edinburgh and London Livingstone
- Tedebrand, Lars-Goran (1972) Sources for the History of Swedish Emigration In Runblom, Harald & Hans Norman (Eds) (1976) *From Sweden to America A History of the Migrations* Minneapolis University of Minnesota Press
- Tepping, Benjamin J (1968) A Model for Optimal Linkage of Records *Journal of the American Statistical Association*, 63 1321-1332
- Thomson, David (1962) *Europe Since Napoleon*, 2nd edition London Longmans
- Trend, J B (1957) *Nations of the Modern World Portugal* London Ernest Benn
- Vasconcellos, Vasco Guedes de (1930) *Compilação de Direito Administrativo Português (de 1832 a 1930) Códigos, Leis, Decretos, Portarias, e Circulares, com índices cronológico e alfabético*, 2 Volumes Lisboa
- Viana, Abel (1953) *Viana do Castelo Escorço Monográfico* Viana do Castelo Aurora do Lima
- Villaverde Cabral, Manuel (1976) *O Desenvolvimento de Capitalismo em Portugal no Século XIX* Lisboa
- Winchester, Ian (1968) *Record Linkage Techniques for Files of Nineteenth Century Historical Records* Paper distributed at the Yale Conference on Nineteenth Century Cities, Yale, Connecticut
- Winchester, Ian (1970) The Association of Historical Records by Man and Computer In Katz, M B (Ed) (1970) *The Hamilton Project*, Vol 2 Ontario Institute for Studies in Education, Toronto
- Winchester, Ian (1973a) On Referring to Ordinary Historical Persons In Wrigley, E Anthony (Ed) (1973) *Identifying People in the Past* London Edward Arnold
- Winchester, Ian (1973b) A Brief Survey of the Algorithmic, Mathematical and Philosophical Literature Relevant to Historical Record Linkage In Wrigley, E Anthony (Ed) (1973) *Identifying People in the Past* London Edward Arnold
- Winchester, Ian (1974) *Partition Principles and the Identification of Historical Individuals/Decision Principles and the Identification of Historical Persons* Unpublished Ph D thesis Oxford University
- Winchester, Ian (1985) *Record Linkage in the Microcomputer Era A Survey* Newsletter No 3 from the Demographic Data Base, University of Umeå, Sweden
- Wrigley, E. Anthony (1966) Family Reconstitution Chapter 4 in *Introduction to Historical Demography* London Methuen
- Wrigley, E Anthony (Ed) (1972) *Nineteenth Century Society* Cambridge University Press
- Wrigley, E Anthony (Ed) (1973) *Identifying People in the Past* London Edward Arnold

- Wrigley, E. Anthony & Roger S Schofield (1973) Nominal Record Linkage by Computer and the Logic of Family Reconstitution In Wrigley, E Anthony (Ed) (1973) *Identifying People in the Past* London Edward Arnold